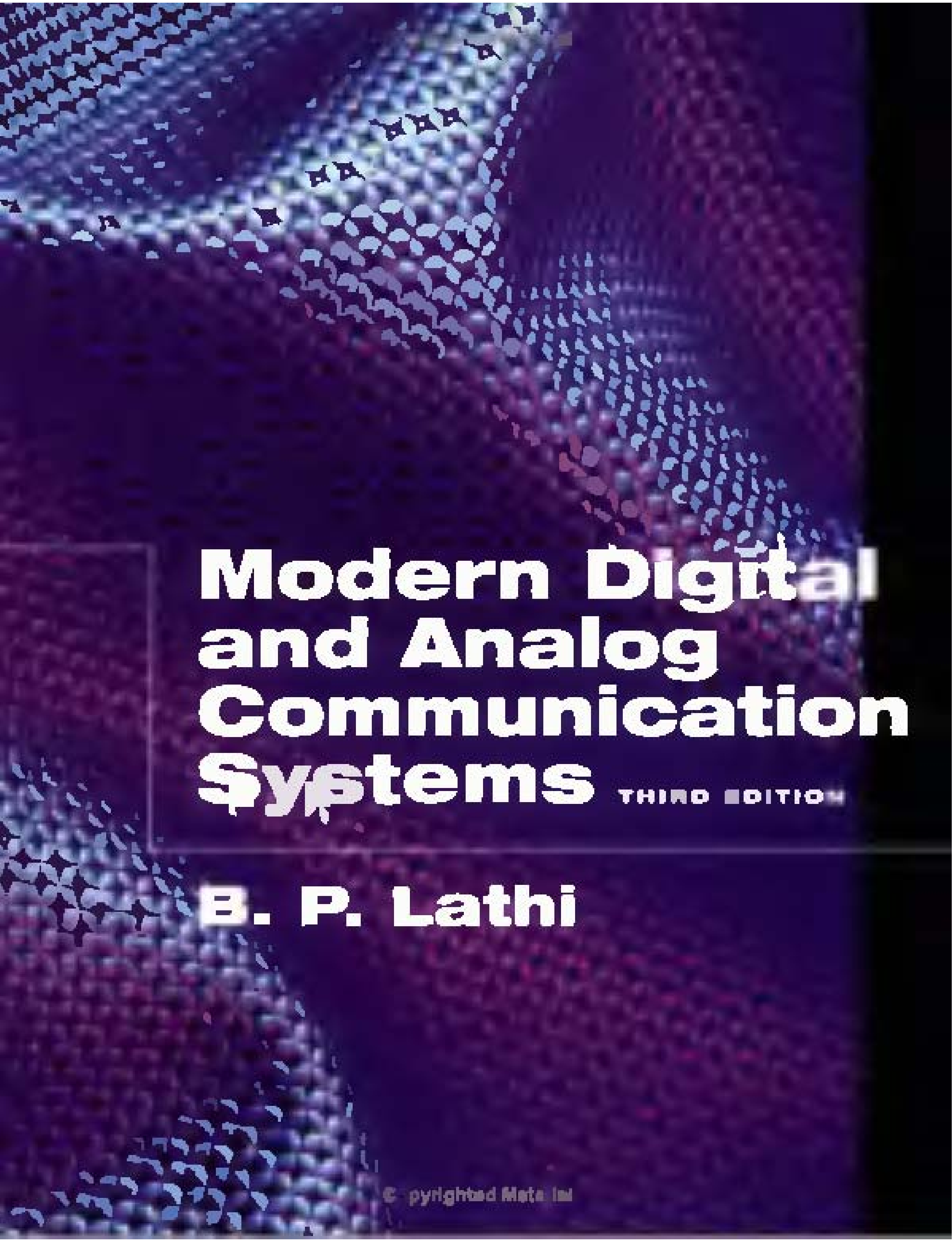


Modern Digital and Analog Communication Systems

THIRD EDITION

B. P. Lathi

Copyrighted Material

SHOUT
SHARE
IT

THE OXFORD SERIES IN ELECTRICAL AND COMPUTER ENGINEERING

SERIES EDITORS

Adel S. Sedra, *Electrical Engineering*
Michael R. Lightner, *Computer Engineering*

SERIES TITLES

Allen and Holberg, *CMOS Analog Circuit Design*
Bobrow, *Elementary Linear Circuit Analysis, 2nd Ed.*
Bobrow, *Fundamentals of Electrical Engineering, 2nd Ed.*
Campbell, *The Science and Engineering of Microelectronic Fabrication*
Chen, *Linear System Theory and Design, 3rd Ed.*
Chen, *System and Signal Analysis, 2nd Ed.*
Comer, *Digital Logic and State Machine Design, 3rd Ed.*
Cooper and McGillem, *Probabilistic Methods of Signal and System Anaylsis, 3rd Ed.*
Franco, *Electric Circuits Fundamentals*
Jones, *Introduction to Optical Fiber Communication Systems*
Krein, *Elements of Power Electronics*
Kuo, *Digital Control Systems, 3rd Ed.*
Lathi, *Modern Digital and Analog Communications Systems, 3rd Ed.*
McGillem and Cooper, *Continuous and Discrete Signal and System Analysis, 3rd Ed.*
Miner, *Lines and Electromagnetic Fields for Engineers*
Roberts, *SPICE, 2nd Ed.*
Santina, Stubberud and Hostetter, *Digital Control System Design, 2nd Ed.*
Schwarz, *Electromagnetics for Engineers*
Schwarz and Oldham, *Electrical Engineering: An Introduction, 2nd Ed.*
Sedra and Smith, *Microelectronic Circuits, 4th Ed.*
Stefani, Savant, and Hostetter, *Design of Feedback Control Systems, 3rd Ed.*
Van Valkenburg, *Analog Filter Design*
Warner and Grung, *Semiconductor Device Electronics*
Wolovich, *Automatic Control Systems*
Yariv, *Optical Electronics in Modern Communications, 5th Ed.*

MODERN DIGITAL AND ANALOG COMMUNICATION SYSTEMS

Third Edition

B. P. LATHI

New York Oxford
OXFORD UNIVERSITY PRESS
1998

Oxford University Press

Oxford New York
Athens Auckland Bangkok Bogota Bombay Buenos Aires
Calcutta Cape Town Dar es Salaam Delhi Florence Hong Kong
Istanbul Karachi Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore Taipei Tokyo Toronto Warsaw
and associated companies in
Berlin Ibadan

Copyright © 1998 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.,
198 Madison Avenue, New York, New York 10016
http://www.oup-usa.org
1-800-334-4249

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Oxford University Press.

Library of Congress Cataloging-in-Publication Data

Lathi, B. P. (Bhagwandas Pannalal)
Modern digital and analog communication systems / B.P. Lathi.—
3rd ed.
p. cm.—(The Oxford series in electrical and computer
engineering)
Includes bibliographical references (p.)
ISBN 0-19-511009-9 (cloth)
1. Telecommunication systems. 2. Digital communications.
3. Statistical communication theory. I. Title II. Series.
TK5101.L333 1998
621.382—dc21 97-16040
CIP

Printing (last digit): 9 8 7 6 5 4 3 2 1
Printed in the United States of American
on acid-free paper

In Memory of
M.E. Van Valkenburg
1921–1997

CONTENTS

PREFACE *xi*

1 INTRODUCTION *1*

Communication System *1*
Analog and Digital Messages *3*
Signal-to-Noise Ratio, Channel Bandwidth, and the Rate of Communication *8*
Modulation *10*
Randomness, Redundancy, and Coding *12*

2 INTRODUCTION TO SIGNALS *14*

2.1 Size of a Signal *14*
2.2 Classification of Signals *20*
2.3 Some Useful Signal Operations *24*
2.4 Unit Impulse Function *28*
2.5 Signals and Vectors *30*
2.6 Signal Comparison: Correlation *35*
2.7 Signal Representation by Orthogonal Signal Set *40*
2.8 Trigonometric Fourier Series *44*
2.9 Exponential Fourier Series *53*
2.10 Numerical Computation of D_n *60*

3 ANALYSIS AND TRANSMISSION OF SIGNALS *71*

3.1 Aperiodic Signal Representation by Fourier Integral *71*
3.2 Transforms of Some Useful Functions *78*
3.3 Some Properties of the Fourier Transform *84*
3.4 Signal Transmission through a Linear System *101*
3.5 Ideal and Practical Filters *106*
3.6 Signal Distortion over a Communication Channel *110*

3.7 Signal Energy and Energy Spectral Density *115*
3.8 Signal Power and Power Spectral Density *123*
3.9 Numerical Computation of Fourier Transform: The DFT *130*

4 AMPLITUDE (LINEAR) MODULATION *151*

4.1 Baseband and Carrier Communication *151*
4.2 Amplitude Modulation: Double Sideband (DSB) *152*
4.3 Amplitude Modulation (AM) *162*
4.4 Quadrature Amplitude Modulation (QAM) *170*
4.5 Amplitude Modulation: Single Sideband (SSB) *171*
4.6 Amplitude Modulation: Vestigial Sideband (VSB) *179*
4.7 Carrier Acquisition *183*
4.8 Superheterodyne AM Receiver *189*
4.9 Television *191*

5 ANGLE (EXPONENTIAL) MODULATION *208*

5.1 Concept of Instantaneous Frequency *208*
5.2 Bandwidth of Angle-Modulated Waves *215*
5.3 Generation of FM Waves *229*
5.4 Demodulation of FM *233*
5.5 Interference in Angle-Modulated Systems *241*
5.6 FM Receiver *245*

6 SAMPLING AND PULSE CODE MODULATION *251*

6.1 Sampling Theorem *251*
6.2 Pulse-Code Modulation (PCM) *262*
6.3 Differential Pulse Code Modulation (DPCM) *278*
6.4 Delta Modulation *281*

7 PRINCIPLES OF DIGITAL DATA TRANSMISSION *294*

7.1 A Digital Communication System *294*
7.2 Line Coding *297*
7.3 Pulse Shaping *310*
7.4 Scrambling *319*
7.5 Regenerative Repeater *322*
7.6 Detection-Error Probability *329*
7.7 *M*-ary Communication *334*
7.8 Digital Carrier Systems *337*
7.9 Digital Multiplexing *342*

8 EMERGING DIGITAL COMMUNICATIONS TECHNOLOGIES *354*

8.1 The North American Hierarchy *354*
8.2 Digital Services *368*
8.3 Broadband Digital Communication: SONET *377*

8.4 Digital Switching Technologies 383
8.5 Broadband Services for Entertainment and Home Office Applications 392
8.6 Video Compression 395
8.7 High-Definition Television (HDTV) 400

9 SOME RECENT DEVELOPMENTS AND MISCELLANEOUS TOPICS 404
9.1 Cellular Telephone (Mobile Radio) System 404
9.2 Spread Spectrum Systems 406
9.3 Transmission Media 416
9.4 Hybrid Circuit: 2-Wire to 4-Wire Conversions 427
9.5 Public Switched Telephone Network 430

10 INTRODUCTION TO THEORY OF PROBABILITY 434
10.1 Concept of Probability 434
10.2 Random Variables 445
10.3 Statistical Averages (Means) 463
10.4 Central-Limit Theorem 472
10.5 Correlation 473
10.6 Linear Mean Square Estimation 476

11 RANDOM PROCESSES 487
11.1 From Random Variable to Random Process 487
11.2 Power Spectral Density of a Random Process 496
11.3 Multiple Random Processes 509
11.4 Transmission of Random Processes through Linear Systems 510
11.5 Bandpass Random Processes 514
11.6 Optimum Filtering: Wiener-Hopf Filter 522

12 BEHAVIOR OF ANALOG SYSTEMS IN THE PRESENCE OF NOISE 532
12.1 Baseband Systems 532
12.2 Amplitude-Modulated Systems 534
12.3 Angle-Modulated Systems 541
12.4 Pulse-Modulated Systems 557
12.5 Optimum Preemphasis-Deemphasis Systems 567

13 BEHAVIOR OF DIGITAL COMMUNICATION SYSTEMS IN THE PRESENCE OF NOISE 577
13.1 Optimum Threshold Detection 577
13.2 General Analysis: Optimum Binary Receiver 582
13.3 Carrier Systems: ASK, FSK, PSK, and DPSK 590
13.4 Performance of Spread Spectrum Systems 601

13.5 *M*-ary Communication *608*
13.6 Synchronization *622*

14 OPTIMUM SIGNAL DETECTION *626*

14.1 Geometrical Representation of Signals: Signal Space *626*
14.2 Gaussian Random Process *632*
14.3 Optimum Receiver *637*
14.4 Equivalent Signal Sets *662*
14.5 Nonwhite (Colored) Channel Noise *669*
14.6 Other Useful Performance Criteria *670*

15 INTRODUCTION TO INFORMATION THEORY *679*

15.1 Measure of Information *679*
15.2 Source Encoding *684*
15.3 Error-Free Communication over a Noisy Channel *690*
15.4 Channel Capacity of a Discrete Memoriless Channel *693*
15.5 Channel Capacity of a Continuous Channel *701*
15.6 Practical Communication Systems in Light of Shannon's Equation *717*

16 ERROR CORRECTING CODES *728*

16.1 Introduction *728*
16.2 Linear Block Codes *731*
16.3 Cyclic Codes *737*
16.4 Burst-Error Detecting and Correcting Codes *745*
16.5 Interlaced Codes for Burst- and Random-Error Correction *746*
16.6 Convolutional Codes *747*
16.7 Comparison of Coded and Uncoded Systems *755*

APPENDIXES *764*

A. Orthogonality of Some Signal Sets *764*
B. Schwarz Inequality *766*
C. Gram-Schmidt Orthogonalization of a Vector Set *768*
D. Miscellaneous *771*

INDEX *775*

PREFACE

The study of communication systems can be divided into two distinct areas:

1. How communication systems work.
2. How they perform in the presence of noise.

The study of each of these two areas, in turn, requires specific tools. To study the first area, the students must be familiar with signal analysis (Fourier techniques), and to study the second area, a basic understanding of probability theory and random processes is essential. For a meaningful comparison of various communication systems, it is necessary to have some understanding of the second area. For this reason many instructors feel that the study of communication systems is not complete unless both of the areas are covered reasonably well. However, it poses one serious problem: the material to be covered is enormous. The two areas along with their tools are overwhelming; it is difficult to cover this material in depth in one course.

The current trend in teaching communication systems is to study the tools in early chapters and then proceed with the study of the two areas of communication. Because too much time is spent in the beginning in studying the tools (without much motivation), there is little time left to study the two proper areas of communication. Consequently, teaching a course in communication systems poses a real dilemma. The second area (statistical aspects) of communication theory is a degree harder than the first area, and it can be properly understood only if the first area is well assimilated. One of the reasons for the dilemma mentioned earlier is our attempt to cover both areas at the same time. The students are forced to grapple with the statistical aspects while also trying to become familiar with how communication systems work. This practice is most unsound pedagogically because it violates the basic fact that one must learn to walk before one can run. The ideal solution would be to offer two courses in sequence, the first course dealing with how communication systems function and the second course dealing with statistical aspects and noise. But in the present curriculum, with so many competing courses, it is difficult to squeeze in two basic courses in the communications area. Some schools do require a course in probability and random processes as a prerequisite. In this case, it is possible to cover both areas reasonably well in a one-semester course. This book, I hope, can be adopted to either case. It can be used as a one-semester survey course in which the deterministic aspects of communication systems are emphasized. It can also be used for a course that deals with deterministic and probabilistic aspects of communication systems. The book provides all the necessary background in probabilities and random processes. However, as stated earlier, it is highly desirable for students to have a good background in probabilities if the course is to be covered in one semester.

The first nine chapters discuss in depth how digital and analog communication systems work, and thus, form a sound, well-rounded, comprehensive survey course in communication systems that is within the reach of an average undergraduate and that can be taught in a one-semester course (about 40 to 45 hours). However, if the students have an adequate background in Fourier analysis and probabilities, it should be possible to cover the first 13 chapters.

Chapter 1 introduces the students to a panoramic view of communication systems. All the important concepts of communication theory are explained qualitatively in a heuristic way. This gets the students deeply interested so that they are encouraged to study the subject. Because of this momentum, they are motivated to study the tool of signal analysis in Chapters 2 and 3, where a student is encouraged to see a signal as a vector, and to think of the Fourier spectrum as a way of representing a signal in terms of its vector components. Chapters 4 and 5 discuss amplitude (linear) and angle (nonlinear) modulation, respectively. Many instructors feel that in this digital age, modulation should be deemphasized with a minimal presence. I feel that modulation is not so much a method of communication as a basic tool of signal processing; it will always be needed not only in the area of communication (digital or analog), but also in many other areas of electrical engineering. Hence, neglecting modulation may prove to be rather shortsighted. Chapter 6 deals with sampling, pulse code modulation (including DPCM), and delta modulation. Chapter 7 discusses transmission of digital data. Some emerging digital technologies in digital data transmission are the subject of Chapter 8. Chapter 9 discusses some recent developments (such as cellular telephone, spread spectrum, global positioning systems), along with miscellaneous topics such as communication media, optical communication, satellite communication, and hybrid circuits. Chapters 10 and 11 provide a reasonably thorough treatment of the theory of probability and random processes. This is the second tool required for the study of communication systems. Every attempt is made to motivate students and sustain their interest through these chapters by providing applications to communications problems wherever possible. Chapters 12 and 13 discuss the behavior of communication systems in the presence of noise. Optimum signal detection is presented in Chapter 14, and information theory is the subject of Chapter 15. Finally, error-control coding is introduced in Chapter 16.

Analog pulse modulation systems such as PAM, PPM, and PWM are deemphasized in comparison to digital schemes (PCM, DPCM, and DM) because the applications of the former in communications are hard to find. In the treatment of angle modulation, rather than compartmentalizing FM and PM, we have provided a generalized treatment of angle modulation, where FM and PM are merely two (of the infinite) special cases. Tone-modulated FM is deemphasized for a sound reason. Since angle modulation is nonlinear, the conclusions derived from tone modulation cannot be blindly applied to modulation by other baseband signals. In fact, these conclusions are misleading in many instances. For example, in the literature PM gets short shrift as being inferior to FM, a conclusion based on tone-modulation analysis.* It is shown in Chapter 12 that PM is, in fact, superior to FM for all practical cases (including audio).

One of the aims in writing this book has been to make learning a pleasant or at least a less intimidating experience for the student by presenting the subject in a clear, understandable, and logically organized manner. Every effort has been made to give an insight—rather than just an understanding—as well as heuristic explanations of theoretical results wherever possible. Many examples are provided for further clarification of abstract results. Even a partial success in achieving my stated goal would make all my toils worthwhile.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the assistance received from several individuals during the preparation of this book. I am greatly indebted to Mr. Maynard Wright, who is a member of several standards committees, for his valuable help in several areas of data transmission. He also contributed Secs. 9.4, 9.5, and part of Sec. 9.3. I greatly appreciate the help of Professor William Jameson from Montana State University, who contributed Chapter 8 (Emerging Digital Communication Technologies). I am much obliged to Prof. Brian Woerner and R.M. Buehrer from Virginia Polytechnic Institute for their contribution. The analysis of spread spectrum systems in Section 13.4 and some parts of Sec. 9.2 are based solely on their contribution. I appreciate the enthusiastic help of Jerry Olup in preparation of the solutions manual. Thanks are also due to several reviewers, especially Profs. W. Green, James Kang, Dan Murphy, W. Jameson, Jeff Reed, R. Vaz, S. Bibyk, C. Alexander and S. Mousavinezhad. I am obliged to Berkeley-Cambridge Press for their permission to use the material (Chapters 2 and 3) from their forthcoming publication *Signal Processing and Linear Systems* by B. P. Lathi. Finally, I owe a debt of gratitude to my wife Rajani for her patience and understanding.

B. P. LATHI

---

* Another reason given for the alleged inferiority of PM is that the phase deviation has to be restricted to a value less than π. It has been shown in Chapter 5 that this is simply not true of band-limited analog signals.

IEEE CODE OF ETHICS

We, the members of the IEEE, in recognition of the importance of our technologies in affecting the quality of life throughout the world, and in accepting a personal obligation to our profession, its members and the communities we serve, do hereby commit ourselves to conduct of the highest ethical and professional manner and agree:

1. to accept responsibility in making engineering decisions consistent with the safety, health, and welfare of the public, and to disclose promptly factors that might endanger the public or the environment;
2. to avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist;
3. to be honest and realistic in stating claims or estimates based on available data;
4. to reject bribery in all of its forms;
5. to improve understanding of technology; its appropriate application, and potential consequences;
6. to maintain and improve our technical competence and to undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations;
7. to seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, and to credit properly the contributions of others;
8. to treat fairly all persons regardless of such factors as race, religion, gender, disability, age, or national origin;
9. to avoid injuring others, their property, reputation, or employment by false or malicious action;
10. to assist colleagues and co-workers in their professional development and to support them in following this code of ethics.

Approved by IEEE Board of Directors, August 1990

For further information please consult the IEEE Ethics Committee WWW page:
http://www.ieee.org/committee.ethics

Copyright © 1997 by IEEE

MODERN DIGITAL AND ANALOG COMMUNICATION SYSTEMS

1 INTRODUCTION

This book examines communication by electrical signals. In the past, messages have been carried by runners, carrier pigeons, drum beats, and torches. These schemes were adequate for the distances and "data rates" of the age. In most parts of the world, these modes of communication have been superseded by electrical communication systems,* which can transmit signals over much longer distances (even to distant planets and galaxies) and at the speed of light.

Electrical communication is reliable and economical; communication technology is alleviating the energy crisis by trading information processing for a more rational use of energy resources. Some examples: Important discussions now mostly communicated face to face in meetings or conferences, often requiring travel, are increasingly using "teleconferring." Similarly, teleshopping and telebanking will provide services by electronic communication, and newspapers may be replaced by electronic news services.

COMMUNICATION SYSTEM

Figure 1.1 shows three examples of communication systems. A typical communication system can be modeled as shown in Fig. 1.2. The components of a communication system are as follows:

The **source** originates a message, such as a human voice, a television picture, a teletype message, or data. If the data is nonelectrical (human voice, teletype message, television picture), it must be converted by an **input transducer** into an electrical waveform referred to as the **baseband signal** or **message signal**.

The **transmitter** modifies the baseband signal for efficient transmission.†

The **channel** is a medium—such as wire, coaxial cable, a waveguide, an optical fiber, or a radio link—through which the transmitter output is sent.

---

* With the exception of the postal service.

† The transmitter consists of one or more of the following subsystems: a preemphasizer, a sampler, a quantizer, a coder, and a modulator. Similarly, the receiver may consist of a demodulator, a decoder, a filter, and a deemphasizer.

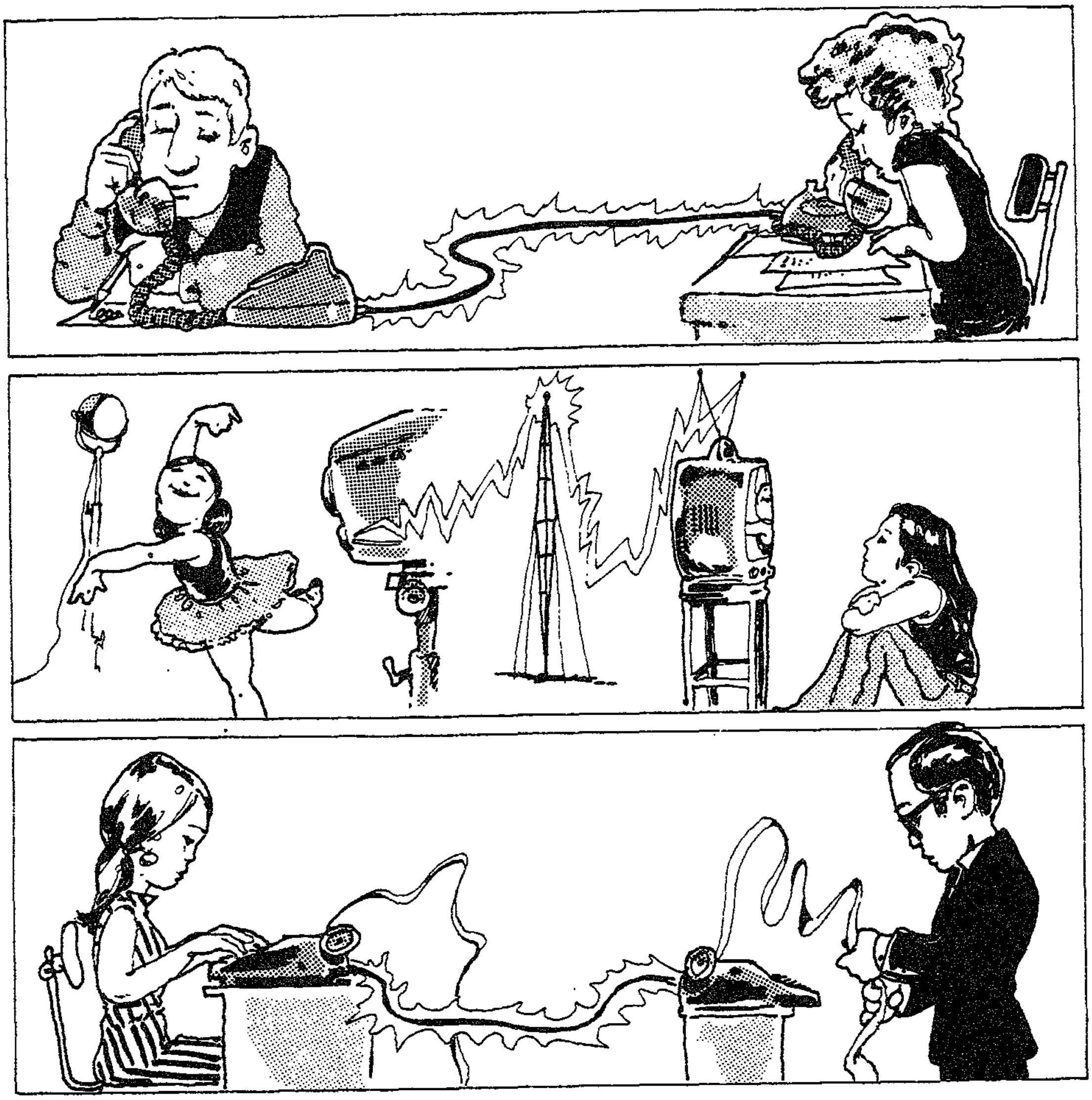


Figure 1.1 Some examples of communications systems.

The **receiver** reprocesses the signal received from the channel by undoing the signal modifications made at the transmitter and the channel. The receiver output is fed to the **output transducer**, which converts the electrical signal to its original form—the message.

The **destination** is the unit to which the message is communicated.

A channel acts partly as a filter to attenuate the signal and distort its waveform. The signal attenuation increases with the length of the channel, varying from a few percent for short distances to orders of magnitude for interplanetary communication. The waveform is distorted because of different amounts of attenuation and phase shift suffered by different frequency components of the signal. For example, a square pulse is rounded or "spread out" during the transmission. This type of distortion, called **linear distortion**, can be partly corrected at the receiver by an equalizer with gain and phase characteristics complementary to those of the channel. The channel may also cause **nonlinear distortion** through attenuation that varies with the signal amplitude. Such distortion can also be partly corrected by a complementary equalizer at the receiver.

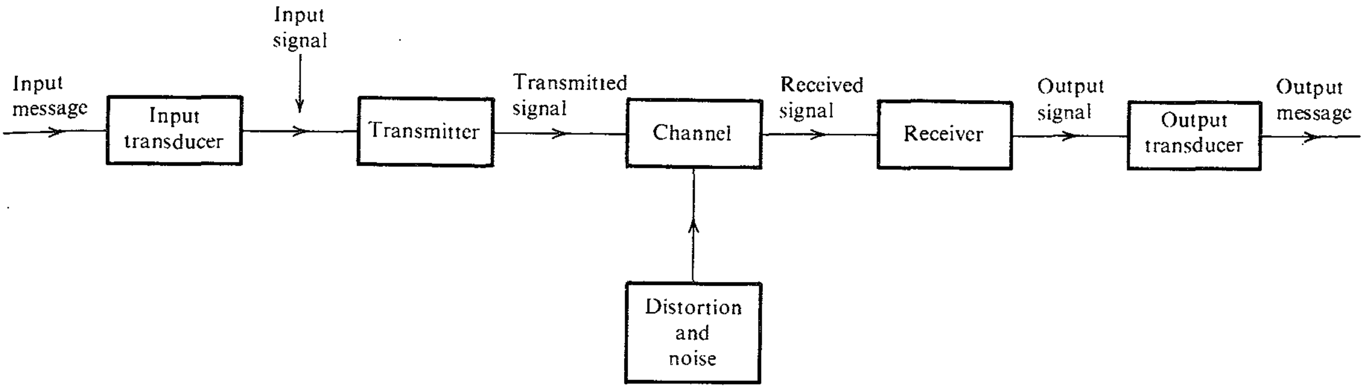


Figure 1.2 Communication system.

The signal is not only distorted by the channel, but it is also contaminated along the path by undesirable signals lumped under the broad term **noise**, which are random and unpredictable signals from causes external and internal. External noise includes interference from signals transmitted on nearby channels, human-made noise generated by faulty contact switches for electrical equipment, automobile ignition radiation, fluorescent lights or natural noise from lightning, as well as electrical storms and solar and intergalactic radiation. With proper care, external noise can be minimized or even eliminated. Internal noise results from thermal motion of electrons in conductors, random emission, and diffusion or recombination of charged carriers in electronic devices. Proper care can reduce the effect of internal noise but can never eliminate it. Noise is one of the basic factors that set limits on the rate of communication.

The **signal-to-noise ratio (SNR)** is defined as the ratio of signal power to noise power. The channel distorts the signal, and noise accumulates along the path. Worse yet, the signal strength decreases while the noise level increases with distance from the transmitter. Thus, the SNR is continuously decreasing along the length of the channel. Amplification of the received signal to make up for the attenuation is of no avail because the noise will be amplified in the same proportion, and the SNR remains, at best, unchanged.*

ANALOG AND DIGITAL MESSAGES

Messages are digital or analog. Digital messages are constructed with a finite number of symbols. For example, printed language consists of 26 letters, 10 numbers, a space, and several punctuation marks. Thus, a text is a digital message constructed from about 50 symbols. Human speech is also a digital message, because it is made up from a finite vocabulary in a language.† Similarly, a Morse-coded telegraph message is a digital message constructed from a set of only **two** symbols—mark and space. It is therefore a **binary** message, implying only two symbols. A digital message constructed with M symbols is called an **M-ary** message.

Analog messages, on the other hand, are characterized by data whose values vary over a continuous range. For example, the temperature or the atmospheric pressure of a certain location can vary over a continuous range and can assume an infinite number of possible values. Similarly, a speech waveform has amplitudes that vary over a continuous range. Over a given time interval, an infinite number of possible different speech waveforms exist, in contrast to only a finite number of possible digital messages.

---

* Actually, amplification further deteriorates the SNR because of the amplifier noise.

† Here we imply the printed text of the speech rather than its details such as the pronunciation of words and varying inflections, pitch, emphasis, and so on. The speech signal from a microphone contains all these details. This signal is an analog signal, and its information content is more than a thousand times the information in the written text of the same speech.

Noise Immunity of Digital Signals

Digital messages are transmitted by using a finite set of electrical waveforms. For example, in the Morse code, a mark can be transmitted by an electrical pulse of amplitude $A/2$, and a space can be transmitted by a pulse of amplitude $-A/2$. In an M-ary case, M distinct electrical pulses (or waveforms) are used; each of the M pulses represents one of the M possible symbols. The task of the receiver is to extract a message from a distorted and noisy signal at the channel output. Message extraction is often easier from digital signals than from analog signals. Consider a binary case: Two symbols are encoded as rectangular pulses of amplitudes $A/2$ and $-A/2$. The only decision at the receiver is the selection between two possible pulses received, not the details of the pulse shape. The decision is readily made with reasonable certainty even if the pulses are distorted and noisy (Fig. 1.3). The digital message in Fig. 1.3a is distorted by the channel, as shown in Fig. 1.3b. Yet, if the distortion is within limits, we can recover the data without error because we need only to make a simple binary decision as to whether the received pulse is positive or negative. Figure 1.3c shows the same data with channel distortion and noise. Here again, the data can be recovered correctly as long as the distortion and the noise are within limits. In contrast, the waveform in an analog message is important, and even a slight distortion or interference in the waveform will cause an error in the received signal. Clearly, a digital communication system is more rugged than an analog communication system in the sense that it can better withstand noise and distortion (as long as they are within a limit).

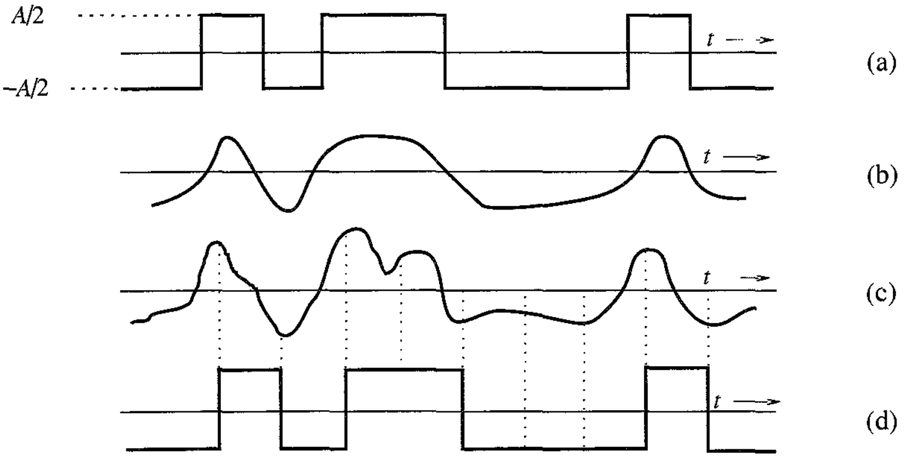


Figure 1.3 (a) Transmitted signal. (b) Received distorted signal (without noise). (c) Received distorted signal (with noise). (d) Regenerated signal (delayed).

Viability of Regenerative Repeaters in Digital Communication

The main reason for the superiority of digital systems over analog ones is the viability of **regenerative repeaters** in the former. Repeater stations are placed along the communication path of a digital system at distances short enough to ensure that noise and distortion remain within a limit. This allows pulse detection with high accuracy. At each repeater station the incoming pulses are detected and new clean pulses are transmitted to the next repeater station. This process prevents the accumulation of noise and distortion along the path by cleaning the pulses periodically at the repeater stations. We can thus transmit messages over longer distances with greater accuracy. For analog systems, there is no way to avoid accumulation of noise and distortion along the path. As a result, the distortion and the noise interference are cumulative over the entire transmission path. To compound the difficulty, the signal is attenuated continuously over the transmission path. Thus, with increasing distance the signal becomes weaker, whereas the distortion and the noise become stronger. Ultimately, the signal, overwhelmed by the distortion and noise, is mutilated. Amplification is of little help, because it enhances the signal and the noise in the same proportion. Consequently, the distance over which an analog message can be transmitted is limited by the transmitter power. Despite these problems, analog communication was used widely and successfully in the past. Because of the advent of optical fiber and the dramatic cost reduction achieved in the fabrication of digital circuitry, almost all new communication systems being installed are digital. But the old analog communication facilities are also in use.

Analog-to-Digital (A/D) Conversion

A meeting ground exists for analog and digital signals: conversion of analog signals to digital signals (A/D conversion). The frequency spectrum of a signal indicates relative magnitudes of various frequency components. The **sampling theorem** (to be proved in Chapter 6) states that if the highest frequency in the signal spectrum is B (in hertz), the signal can be reconstructed from its samples, taken at a rate not less than $2B$ samples per second. This means that in order to transmit the information in a continuous-time signal, we need only transmit its samples (Fig. 1.4). Unfortunately, the sample values are still not digital because they lie in a continuous range and can take on any one of the infinite values in the range. We are back where we started! This difficulty is neatly resolved by what is known as **quantization**, where each sample is approximated, or "rounded off," to the nearest quantized level, as shown in Fig. 1.4. Amplitudes of the signal $m(t)$ lie in the range $(-m_p,\ m_p)$, which is partitioned into L intervals, each of magnitude $\Delta v = 2m_p/L$. Each sample amplitude is approximated to the midpoint of the interval in which the sample value falls. Each sample is now approximated to one of the L numbers. The information is thus digitized.

The quantized signal is an approximation of the original one. We can improve the accuracy of the quantized signal to any desired degree by increasing the number of levels L.

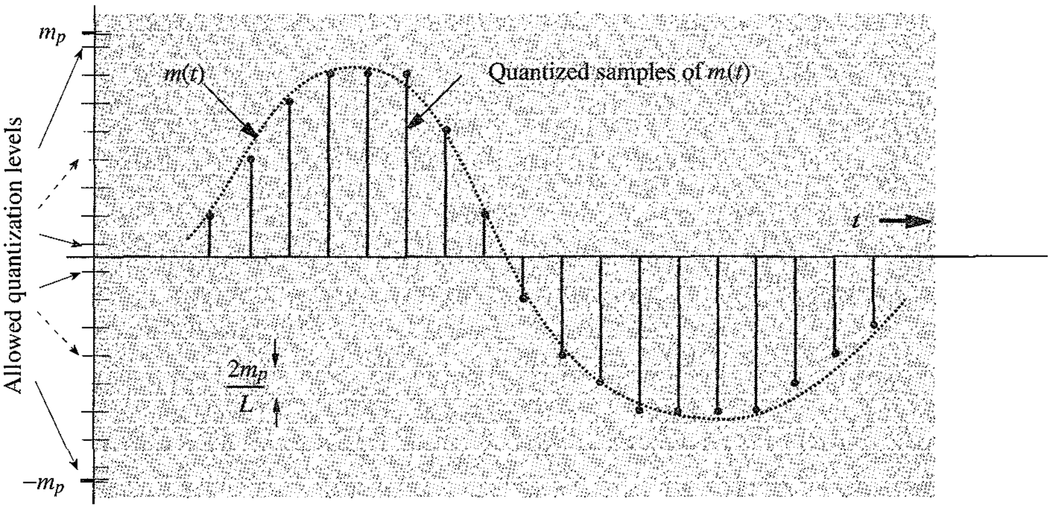


Figure 1.4 Analog-to-digital conversion of a signal.

For intelligibility of voice signals, for example, $L = 8$ or 16 is sufficient. For commercial use, $L = 32$ is a minimum, and for telephone communication, $L = 128$ or 256 is commonly used.

During each sampling interval, we transmit one quantized sample, which takes on one of the L values. This requires L distinct waveforms, which may be constructed, for example, by using a basic rectangular pulse of amplitude $A/2$ and its multiples (for instance, $\pm A/2, \pm 3A/2, \pm 5A/2, \dots, \pm[(L-1)A/2]$, as shown in Fig. 1.5) to form L distinct waveforms to be assigned to the L values to be transmitted. Amplitudes of any two of these waveforms are separated by at least A to guard against noise interference and channel distortion. Another possibility is to use fewer than L waveforms and form their combinations (codes) to yield L distinct patterns. As an example, for the case $L = 16$ we may use 16 pulses ($\pm A/2, \pm 3A/2, \dots, \pm 15A/2$, as shown in Fig. 1.5). The second alternative is to use combinations of only two basic pulses, $A/2$ and $-A/2$. A sequence of four such pulses gives $2 \times 2 \times 2 \times 2 = 16$ distinct patterns, as shown in Fig. 1.6. We can assign one pattern to each of the 16 quantized values to be transmitted. Each quantized sample is now coded into a sequence of four binary pulses. This is the so-called binary case, where signaling is carried out by means of only two basic pulses (or symbols).*

The binary case is of great practical importance because of its simplicity and ease of detection. Virtually all digital communication today is binary. This scheme of transmitting data by digitizing and then using pulse codes to transmit the digitized data is known as **pulse-code modulation (PCM)**.

A typical distorted binary signal with noise acquired over the channel is shown in Fig. 1.3. If A is sufficiently large compared to typical noise amplitudes, the receiver can still correctly distinguish between the two pulses. The pulse amplitude is typically 5 to 10 times the rms noise amplitude. For such a high SNR, the probability of error at the receiver is less than 10^{-6}; that is, on the average, the receiver will make less than one error per million pulses. The effect

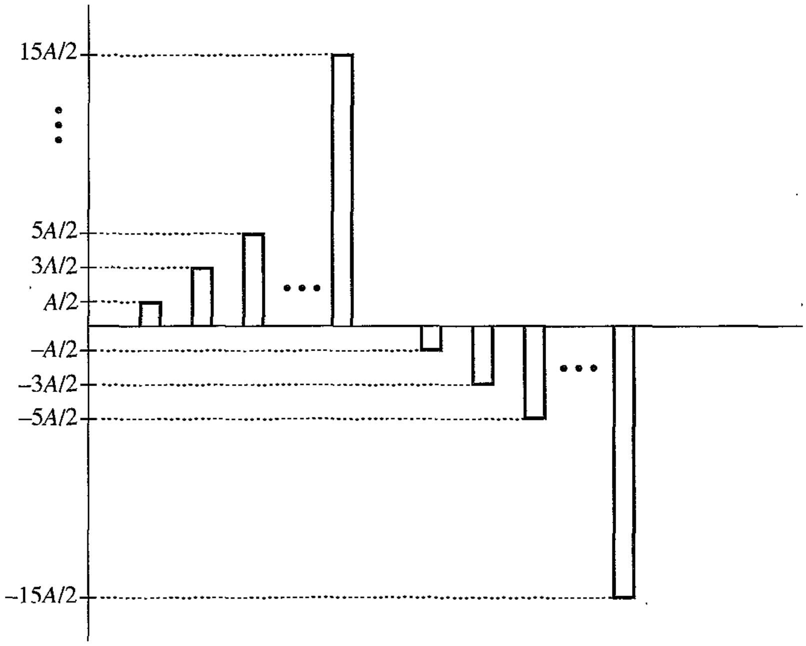


Figure 1.5 Multiamplitude pulse code that uses L amplitude levels.

---

* An intermediate case exists where we use four basic pulses (quaternary pulses) of amplitudes $\pm A/2$ and $\pm 3A/2$. A sequence of two quaternary pulses can form $4 \times 4 = 16$ distinct levels of values.

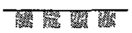
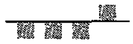
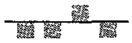
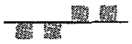
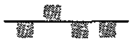
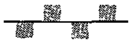
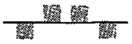
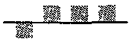
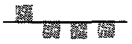
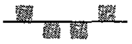
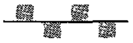
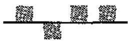
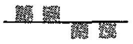
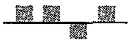
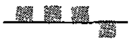
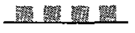
Digit	Binary equivalent	Pulse code waveform
0	0000	
1	0001	
2	0010	
3	0011	
4	0100	
5	0101	
6	0110	
7	0111	
8	1000	
9	1001	
10	1010	
11	1011	
12	1100	
13	1101	
14	1110	
15	1111	

Figure 1.6 Example of a binary pulse code.

of random channel noise and distortion is thus practically eliminated. Hence, when analog signals are transmitted by digital means, the only error, or uncertainty, in the received signal is that caused by quantization. By increasing L, we can reduce the uncertainty, or error, caused by quantization to any desired amount. At the same time, because of the use of regenerative repeaters, we can transmit signals over a much longer distance than would have been possible for the analog signal. As will be seen later in this chapter, the price for all these benefits of digital communication is paid in terms of increased bandwidth of transmission.

From all this discussion, we arrive at a rather interesting (and by no means obvious) conclusion—that every possible communication can be carried on with a minimum of two symbols. Thus, merely by using a proper sequence of a wink of the eye, one can convey any message, be it a conversation, a book, a movie, or an opera star's singing. Every possible detail (such as various shades of colors of the objects and tones of the voice, etc.) that is reproducible on a movie screen or on the best quality color television can be conveyed with no less accuracy, merely by a wink of an eye.*

Although PCM was invented by P. M. Rainey in 1926 and rediscovered by A. H. Reeves in 1939, it was not until the early sixties that Bell Laboratories installed the first communication link using PCM. The cost and size of vacuum tube circuits were the chief impediments to the use of PCM in the early days. It was the transistor that made PCM practicable.

---

* Of course, to convey the information in a movie or a television program in real time, the winking would have to be at an inhumanly high speed.

SIGNAL-TO-NOISE RATIO, CHANNEL BANDWIDTH, AND THE RATE OF COMMUNICATION

The fundamental parameters that control the rate and quality of information transmission are the channel bandwidth B and the signal power S. The appropriate quantitative relationships will be derived later. Here we shall demonstrate these relationships qualitatively.

The **bandwidth** of a channel is the range of frequencies that it can transmit with reasonable fidelity. For example, if a channel can transmit with reasonable fidelity a signal whose frequency components occupy a range from 0 (dc) up to a maximum of 5000 Hz (5 kHz), the channel bandwidth B is 5 kHz.

To understand the role of B, consider the possibility of increasing the speed of information transmission by time compression of the signal. If a signal is compressed in time by a factor of 2, it can be transmitted in half the time, and the speed of transmission is doubled. Compression by a factor of 2, however, causes the signal to "wiggle" twice as fast, implying that the frequencies of its components are doubled. To transmit this compressed signal without distortion, the channel bandwidth must also be doubled. Thus, the rate of information transmission is directly proportional to B. More generally, if a channel of bandwidth B can transmit N pulses per second, then to transmit KN pulses per second we need a channel of bandwidth KB. To reiterate, the number of pulses per second that can be transmitted over a channel is directly proportional to its bandwidth B.

The **signal power** S plays a dual role in information transmission. First, S is related to the quality of transmission. Increasing S, the signal power, reduces the effect of channel noise, and the information is received more accurately, or with less uncertainty. A larger signal-to-noise ratio (SNR) also allows transmission over a longer distance. In any event, a certain minimum SNR is necessary for communication.

The second role of the signal power is not as obvious, although it is very important. We shall demonstrate that the channel bandwidth B and the signal power S are exchangeable; that is, to maintain a given rate and accuracy of information transmission, we can trade S for B, and vice versa. Thus, one may reduce B if one is willing to increase S, or one may reduce S if one is willing to increase B. The rigorous proof of this will be provided in Chapter 15. Here we shall give only a "plausibility argument."

Consider the PCM scheme discussed earlier, with 16 quantization levels ($L = 16$). Here we may use 16 distinct pulses of amplitudes $\pm A/2, \pm 3A/2, \ldots, \pm 15A/2$ to represent the 16 levels (a 16-ary case). Each sample is transmitted by one of the 16 pulses during the sampling interval T_s. The amplitudes of these pulses range from $-15A/2$ to $15A/2$. Alternately, we may use the binary scheme, where a group of four binary pulses is used to transmit each sample during the sampling interval T_s. In the latter case, the transmitted power is reduced considerably because the peak amplitude of transmitted pulses is only $A/2$, as compared to the peak amplitude $15A/2$ in the 16-ary case. In the binary case, however, we need to transmit four pulses in each interval T_s instead of just one pulse required in the 16-ary case. Thus, the required channel bandwidth in the binary case is 4 times as great as that for the 16-ary case. Despite the fact that the binary case requires 4 times as many pulses, its power is less than the power required for the 16-ary case by a factor of $255/12 = 21.25$, as shown later in Eq. 13.51a.[*] In both cases, the minimum amplitude separation between transmitted pulses is A, and we therefore have about the same error probability at the receiver.[†] This means the quality of the received signal is about the same in both cases. In the binary case, the transmitted signal power is reduced at the cost of increased bandwidth. We have demonstrated here the exchangeability of S with B. Later we shall see that relatively little increase in B enables a significant reduction in S.

In conclusion, the two primary communication resources are the bandwidth and the transmitted power. In a given communication channel, one resource may be more valuable than the other, and the communication scheme should be designed accordingly. A typical telephone channel, for example, has a limited bandwidth (3 kHz), but a lot of power is available. On the other hand, in space vehicles, infinite bandwidth is available but the power is limited. Hence, the communication schemes required in the two cases are radically different.

Since the SNR is proportional to the power S, we can say that SNR and bandwidth are exchangeable. It will be shown in Chapter 15 that the relationship between the bandwidth expansion factor and the SNR is exponential. Thus, if a given rate of information transmission requires a channel bandwidth B_1 and a signal-to-noise ratio SNR_1, then it is possible to transmit the same information over a channel bandwidth B_2 and a signal-to-noise ratio SNR_2, where

$$\text{SNR}_2 \simeq \text{SNR}_1^{B_1/B_2} \tag{1.1}$$

Thus, if we double the channel bandwidth, the required SNR is just a square root of the former SNR, and tripling the channel bandwidth reduces the corresponding SNR to just a cube root of the former SNR. Therefore, a relatively small increase in channel bandwidth buys a large advantage in terms of reduced transmission power. But a large increase in transmitted power buys a meager advantage in bandwidth reduction. Hence, in practice, the exchange between B and SNR is usually in the sense of increasing B to reduce transmitted power, and rarely the other way around.

Equation (1.1) gives the upper bound on the exchange between SNR and B. Not all systems are capable of achieving this bound. For example, frequency modulation (FM) is one scheme that is commonly used in radio broadcasting for improving the signal quality at the receiver by increasing the transmission bandwidth. We shall see that an FM system does not make efficient use of bandwidth in reducing the required SNR, and its performance falls far short of that in Eq. (1.1). PCM, on the other hand, comes close (within 10 dB) to realizing the performance in Eq. (1.1). Generally speaking, the transmission of signals in digital form comes much closer to the realization of the limit in Eq. (1.1) than does the transmission of signals in analog form.

The limitation imposed on communication by the channel bandwidth and the SNR is dramatically highlighted by Shannon's equation,[‡]

$$C = B \log_2(1 + \text{SNR}) \quad \text{bit/s} \tag{1.2}$$

---

[*] To explain this behavior qualitatively, let the number of symbols used be M ($M = 16$ in the present case) instead of 2 (binary case). We shall see later that the power of a pulse is proportional to its amplitude. Hence, the signal power increases as $(M - 1)^2$, but n, the number of binary pulses per sample, increases only as the logarithm of M.

[†] Not quite true! We use this approximation to keep our argument simple and nonquantitative at this point.

[‡] This is true for a certain kind of noise—white gaussian noise.

Here C is the rate of information transmission per second. This rate C (known as the channel capacity) is the maximum number of binary symbols (bits) that can be transmitted per second with a probability of error arbitrarily close to zero. In other words, a channel can transmit $B \log_2 (1 + \text{SNR})$ binary digits, or symbols, per second as accurately as one desires. Moreover, it is impossible to transmit at a rate higher than this without incurring errors. Shannon's equation clearly brings out the limitation on the rate of communication imposed by B and SNR. If there were no noise on the channel ($N = 0$), $C = \infty$, and communication would cease to be a problem. We could then transmit any amount of information in the world over a channel. This can be readily verified. If noise were zero, there would be no uncertainty in the received pulse amplitude, and the receiver would be able to detect any pulse amplitude without ambiguity. The minimum pulse-amplitude separation A can be arbitrarily small, and for any given pulse, we have an infinite number of levels available. We can assign one level to every possible message. For example, the contents of this book will be assigned one level; if it is desired to transmit this book, all that is needed is to transmit one pulse of that level. Because an infinite number of levels are available, it is possible to assign one level to any conceivable message. Cataloging of such a code may not be practical, but that is beside the point. The point is that if the noise is zero, communication ceases to be a problem, at least theoretically. Implementation of such a scheme would be difficult because of the requirement of generation and detection of pulses of precise amplitudes. Such practical difficulties would then set a limit on the rate of communication.

In conclusion, we have demonstrated qualitatively the basic role played by B and SNR in limiting the performance of a communication system. These two parameters then represent the ultimate limitation on a rate of communication. We have also demonstrated the possibility of trade or exchange between these two basic parameters.

Equation (1.1) can be derived from Eq. (1.2). It should be remembered that Shannon's result represents the upper limit on the rate of communication over a channel and can be achieved only with a system of monstrous and impractical complexity, and with a time delay in reception approaching infinity. Practical systems operate at rates below the Shannon rate. In Chapter 15, we shall derive Shannon's result and compare the efficiencies of various communication systems.

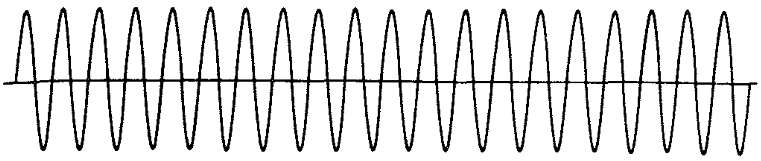
MODULATION

Baseband signals produced by various information sources are not always suitable for direct transmission over a given channel. These signals are usually further modified to facilitate transmission. This conversion process is known as **modulation**. In this process, the baseband signal is used to modify some parameter of a high-frequency carrier signal.

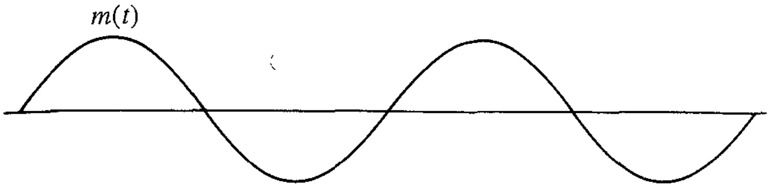
A **carrier** is a sinusoid of high frequency, and one of its parameters—such as amplitude, frequency, or phase—is varied in proportion to the baseband signal $m(t)$. Accordingly, we have amplitude modulation (AM), frequency modulation (FM), or phase modulation (PM). Figure 1.7 shows a baseband signal $m(t)$ and the corresponding AM and FM waveforms. In AM, the carrier amplitude varies in proportion to $m(t)$, and in FM, the carrier frequency varies in proportion $m(t)$.

At the receiver, the modulated signal must pass through a reverse process called **demodulation** in order to reconstruct the baseband signal.

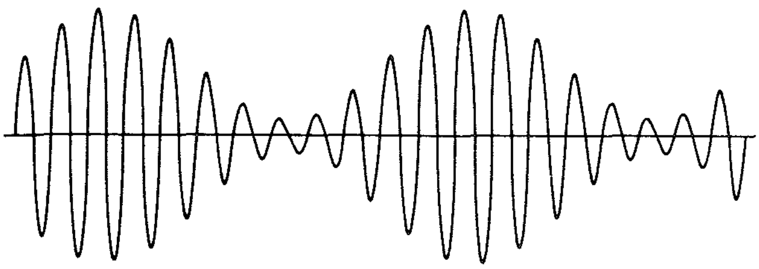
Figure 1.7 Modulation.



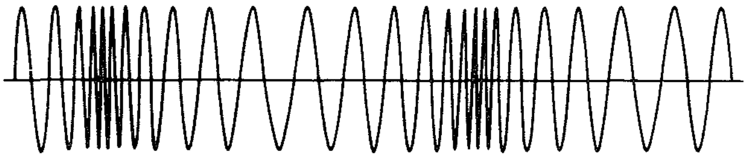
Carrier (a)



Modulating (baseband) signal (b)



Amplitude-modulated wave (c)



Frequency-modulated wave (d)

As mentioned earlier, modulation is used to facilitate transmission. Some of the important reasons for modulation are given next.

Ease of Radiation

For efficient radiation of electromagnetic energy, the radiating antenna should be on the order of one-tenth or more of the wavelength of the signal radiated. For many baseband signals, the wavelengths are too large for reasonable antenna dimensions. For example, the power in a speech signal is concentrated at frequencies in the range of 100 to 3000 Hz. The corresponding wavelength is 100 to 3000 km. This long wavelength would necessitate an impracticably large antenna. Instead, we modulate a high-frequency carrier, thus translating the signal spectrum to the region of carrier frequencies that corresponds to a much smaller wavelength. For example, a 1-MHz carrier has a wavelength of only 300 m and requires an antenna whose size is on the order of 30 m. In this aspect, modulation is like letting the baseband signal hitchhike on a high-frequency sinusoid (carrier). The carrier and the baseband signal may be compared to a stone and a piece of paper. If we wish to throw a piece of paper, it cannot go too far by itself. But by wrapping it around a stone (a carrier), it can be thrown over a longer distance.

Simultaneous Transmission of Several Signals

Consider the case of several radio stations broadcasting audio baseband signals directly, without any modification. They would interfere with each other because the spectra of all the signals occupy more or less the same bandwidth. Thus, it would be possible to broadcast from only one radio or television station at a time. This is wasteful because the channel bandwidth may be much larger than that of the signal. One way to solve this problem is to use modulation. We can use various audio signals to modulate different carrier frequencies, thus translating each signal to a different frequency range. If the various carriers are chosen sufficiently far apart in frequency, the spectra of the modulated signals will not overlap and thus will not interfere with each other. At the receiver, one can use a tunable bandpass filter to select the desired station or signal. This method of transmitting several signals simultaneously is known as **frequency-division multiplexing (FDM)**. Here the bandwidth of the channel is shared by various signals without any overlapping.

Another method of multiplexing several signals is known as **time-division multiplexing (TDM)**. This method is suitable when a signal is in the form of a pulse train (as in PCM). The pulses are made narrower, and the spaces that are left between pulses are used for pulses from other signals. Thus, in effect, the transmission time is shared by a number of signals by interleaving the pulse trains of various signals in a specified order. At the receiver, the pulse trains corresponding to various signals are separated.

Effecting the Exchange of SNR with *B*

We have shown earlier that it is possible to exchange SNR with the bandwidth of transmission. FM or PM can effect such an exchange. The amount of modulation (to be defined later) used controls the exchange of SNR and the transmission bandwidth.

RANDOMNESS, REDUNDANCY, AND CODING

Randomness plays an important role in communication. As noted earlier, one of the limiting factors in the rate of communication is noise, which is a random signal. Randomness is also closely associated with information. Indeed, randomness is the essence of communication. Randomness means unpredictability, or uncertainty, of the outcome. If a source had no unpredictability, or uncertainty, it would be known beforehand and would convey no information. Probability is the measure of certainty, and information is associated with probability. If a person winks, it conveys some information in a given context. But if a person were to wink continuously with the regularity of a clock, it would convey no meaning. The unpredictability of the winking is what gives the information to the signal. What is more interesting, however, is that from the engineering point of view, also, information is associated with uncertainty. The information of a message, from the engineering point of view, is defined as a quantity proportional to the minimum time needed to transmit it. Consider the Morse code, for example. In this code, various combinations of marks and spaces (code words) are assigned to each letter. In order to minimize the transmission time, shorter code words are assigned to more frequently occurring (more probable) letters (such as e, t, and a) and longer code words are assigned to rarely occurring (less probable) letters (such as x, q, and z). Thus, the time required to transmit a message is closely related to the probability of its occurrence. It will be shown in Chapter 15 that for digital signals, the overall transmission time is minimized if a message (or symbol) of probability P is assigned a code word with a length proportional to $\log(1/P)$. Hence, from an engineering point of view, the information of a message with probability P is proportional to $\log(1/P)$.

Redundancy also plays an important role in communication. It is essential for reliable communication. Because of redundancy, we are able to decode a message accurately despite errors in the received message. Redundancy thus helps combat noise. All languages are redundant. For example, English is about 50 percent redundant; that is, on the average, we may throw out half of the letters or words without destroying the message. This also means that in any English message, the speaker or the writer has free choice over half the letters or words, on the average. The remaining half is determined by the statistical structure of the language. If all the redundancy of English were removed, it would take about half the time to transmit a telegram or telephone conversation. If an error occurs at the receiver, however, it would be rather difficult to make sense out of the received message. The redundancy in a message, therefore, plays a useful role in combating the noise in the channel. This same principle of redundancy applies in coding messages. A deliberate redundancy is used to combat the noise. For example, in order to transmit samples with $L = 16$ quantizing levels, we may use a group of four binary pulses, as shown in Fig. 1.6. In this coding scheme, no redundancy exists. If an error occurs in the reception of even one of the pulses, the receiver will produce a wrong value. Here we may use redundancy to eliminate the effect of possible errors caused by channel noise or imperfections. Thus, if we add to each code word one more pulse of such polarity as to make the number of positive pulses even, we have a code that can detect a single error in any place. Thus, to the code words **0001** we add a fifth pulse, of positive polarity, to make a new code word, **00011**. Now the number of positive pulses is 2 (even). If a single error occurs in any position, this parity will be violated. The receiver knows that an error has been made and can request retransmission of the message. This is a very simple coding scheme. It can only detect an error, but cannot locate it. Moreover, it cannot detect an even number of errors. By introducing more redundancy, it is possible not only to detect but also to correct errors. For example, for $L = 16$, it can be shown that properly adding three pulses will not only detect but also correct a single error occurring at any location. This subject of error-correcting codes will be discussed in Chapter 16.

2

INTRODUCTION TO SIGNALS

In this chapter we discuss certain basic signal concepts. Signals are processed by systems. We shall start with explaining the terms *signals* and *systems*.

Signals

A **signal**, as the term implies, is a set of information or data. Examples include a telephone or a television signal, monthly sales of a corporation, or the daily closing prices of a stock market (e.g., the Dow Jones averages). In all these examples, the signals are functions of the independent variable *time*. This is not always the case, however. When an electrical charge is distributed over a surface, for instance, the signal is the charge density, a function of *space* rather than time. In this book we deal almost exclusively with signals that are functions of time. The discussion, however, applies equally well to other independent variables.

Systems

Signals may be processed further by **systems**, which may modify them or extract additional information from them. For example, an antiaircraft gun operator may want to know the future location of a hostile moving target, which is being tracked by a radar. Knowing the radar signal, the antiaircraft gun operator knows the past location and velocity of the target. By properly processing the radar signal (the input), we can approximately estimate the future location of the target. Thus, a system is an entity that *processes* a set of signals (**inputs**) to yield another set of signals (**outputs**). A system may be made up of physical components, as in electrical, mechanical, or hydraulic systems (hardware realization), or it may be an algorithm that computes an output from an input signal (software realization).

2.1 SIZE OF A SIGNAL

The size of any entity is a number that indicates the largeness or strength of that entity. Generally speaking, the signal amplitude varies with time. How can a signal that exists over a certain time interval with varying amplitude be measured by one number that will indicate the signal size or signal strength? Such a measure must consider not only the signal amplitude, but also its duration. For instance, if we are to devise a single number V as a measure of the size of a human being, we must consider not only his or her width (girth), but also the height. The product of girth and height is a reasonable measure of the size of a person. If we wish to be a little more precise, we should average this product over the entire length of the person. If we make the simplifying assumption that the shape of a person is a cylinder of radius r, which varies with the height h of the person, then a reasonable measure of the size of a person of height H is the person's volume V, given by

$$V = \pi \int_0^H r^2(h)\, dh$$

Signal Energy

Arguing in this manner, we may consider the area under a signal $g(t)$ as a possible measure of its size, because it takes account of not only the amplitude, but also the duration. However, this will be a defective measure because $g(t)$ could be a large signal, yet its positive and negative areas could cancel each other, indicating a signal of small size. This difficulty can be corrected by defining the signal size as the area under $g^2(t)$, which is always positive. We call this measure the **signal energy** E_g, defined (for a real signals) as

$$E_g = \int_{-\infty}^{\infty} g^2(t)\, dt \tag{2.1}$$

This definition can be generalized to a complex valued signal $g(t)$ as

$$E_g = \int_{-\infty}^{\infty} |g(t)|^2\, dt \tag{2.2}$$

There are also other possible measures of signal size, such as the area under $|g(t)|$. The above energy measure, however, is not only more tractable mathematically, but is also more meaningful (as shown later) in the sense that it is indicative of the energy that can be extracted from the signal.

Signal Power

The signal energy must be finite for it to be a meaningful measure of the signal size. A necessary condition for the energy to be finite is that the signal amplitude $\to 0$ as $|t| \to \infty$ (Fig. 2.1a). Otherwise the integral in Eq. (2.1) will not converge.

If the amplitude of $g(t)$ does not $\to 0$ as $|t| \to \infty$ (Fig. 2.1b), the signal energy is infinite. A more meaningful measure of the signal size in such a case would be the time average of the energy (if it exists), which is the average power P_g defined (for a real signal) by

$$P_g = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t)\, dt \tag{2.3}$$

We can generalize this definition for a complex signal $g(t)$ as

$$P_g = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |g(t)|^2\, dt \tag{2.4}$$

Observe that the signal power P_g is the time average (mean) of the signal amplitude squared, that is the **mean-squared** value of $g(t)$. Indeed, the square root of P_g is the familiar **root mean square (rms)** value of $g(t)$.

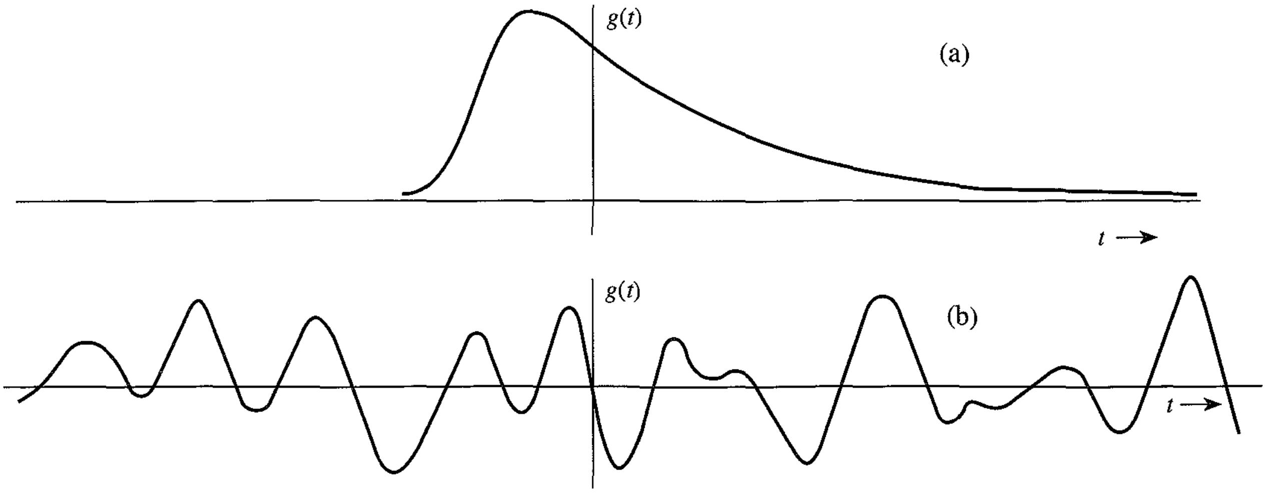


Figure 2.1 Examples of signals. (a) Signal with finite energy. (b) Signal with finite power.

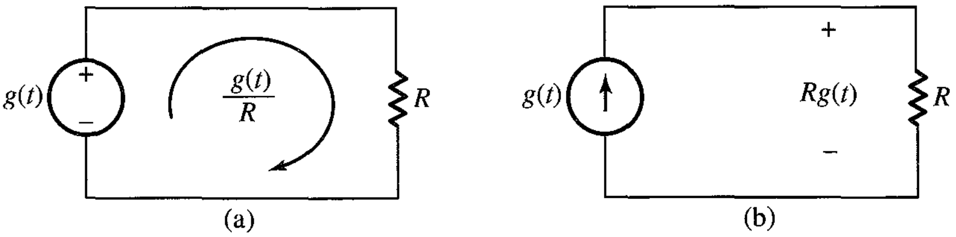
The mean of an entity averaged over a large time interval approaching infinity exists if the entity is either periodic or has a statistical regularity. If such a condition is not satisfied, the average may not exist. For instance, a ramp signal $g(t) = t$ increases indefinitely as $|t| \to \infty$, and neither the energy nor the power exists for this signal.

Comments

The signal energy as defined in Eq. (2.1) or Eq. (2.2) does not indicate the actual energy of the signal because the signal energy depends not only on the signal, but also on the load. It can, however, be interpreted as the energy dissipated in a normalized load of a 1-ohm resistor. If a voltage $g(t)$ is applied across an R-ohm resistor, the current through the resistor is $g(t)/R$ (Fig. 2.2a), and the instantaneous power dissipated would be $v(t)i(t) = g^2(t)/R$. The energy dissipated, being the integral of the instantaneous power, is

$$\text{Energy dissipated} = \int_{-\infty}^{\infty} \frac{g^2(t)}{R}\,dt = \frac{E_g}{R} \tag{2.5}$$

If $R = 1$, the energy dissipated in the resistor is E_g. Thus, the signal energy E_g could be interpreted as the energy dissipated in a unit resistor if a voltage $g(t)$ were applied across this unit resistor. From Fig. 2.2b, it follows that E_g may also be interpreted as the energy dissipated in a unit resistor if a current $g(t)$ were passed through this unit resistor. Parallel observation applies to signal power as defined in Eq. (2.3) or Eq. (2.4).



(a) (b)

Figure 2.2 Computation of the actual energy dissipated across a load.

The measure of "energy" (or "power") is therefore indicative of the energy (or power) capability of the signal, and not the actual energy. For this reason the concepts of conservation of energy should not be applied to the measure of signal energy. These measures are but convenient indicators of the signal size. For instance, if we approximate a signal $g(t)$ by another signal $z(t)$, the error in the approximation is $e(t) = g(t) - z(t)$. The energy (or power) of $e(t)$ is a convenient indicator of the goodness of the approximation. It provides us with a quantitative measure of determining the closeness of the approximation. It also allows us to determine if one approximation is better than the other. In communication systems, during transmission over a channel, message signals are corrupted by unwanted signals (noise). The quality of the received signal is judged by the relative sizes of the desired signal and the unwanted signal (noise). In this case the ratio of the message signal and the noise signal powers (SNR) is a good indication of the received signal quality.

Units of Energy and Power: Equations (2.1) and (2.2) are not correct dimensionally. This is because here we are using the term *energy* not in its conventional sense, but to indicate the signal size. The same observations apply to Eqs. (2.3) and (2.4) for power. In the present context the units of energy and power depend on the nature of the signal $g(t)$. If $g(t)$ is a voltage signal, its energy E_g has units of volts squared seconds, and its power P_g has units of volts squared. If $g(t)$ is a current signal, these units will be amperes squared seconds, and amperes squared, respectively.

EXAMPLE 2.1 Determine the suitable measures of the signals in Fig. 2.3.

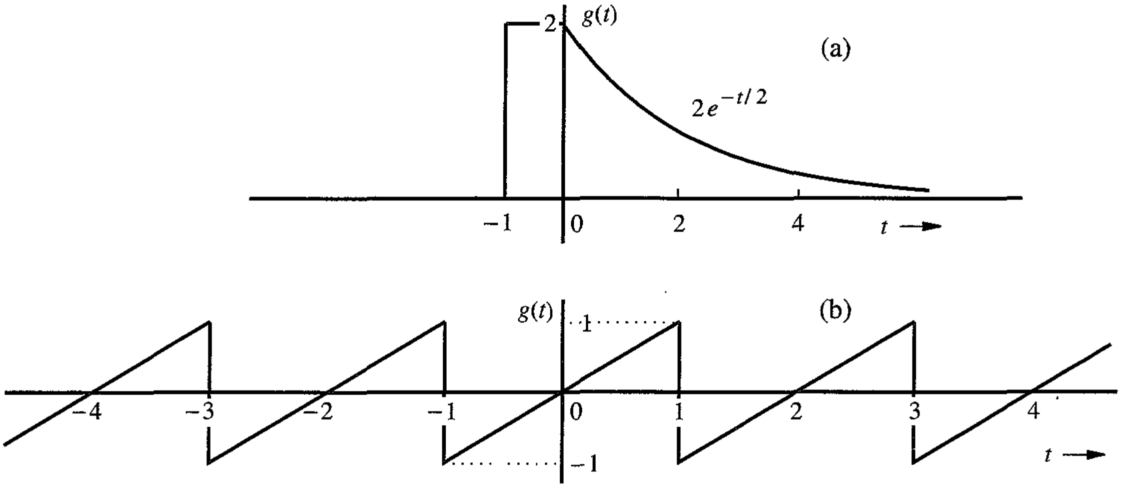


Figure 2.3 Signal for Example 2.1.

The signal in Fig. 2.3a → 0 as $|t| \to \infty$. Therefore, the suitable measure for this signal is its energy E_g given by

$$E_g = \int_{-\infty}^{\infty} g^2(t)\, dt = \int_{-1}^{0} (2)^2\, dt + \int_{0}^{\infty} 4e^{-t}\, dt = 4 + 4 = 8$$

The signal in Fig. 2.3b does not → 0 as $|t| \to \infty$. However, it is periodic, and therefore its power exists. We can use Eq. (2.3) to determine its power. We can simplify the procedure for periodic signals by observing that a periodic signal repeats regularly each period (2 seconds in this case). Therefore, averaging $g^2(t)$ over an infinitely large interval is identical to averaging it over one period (2 seconds in this case). Thus,

$$P_g = \frac{1}{2}\int_{-1}^{1} g^2(t)\,dt = \frac{1}{2}\int_{-1}^{1} t^2\,dt = \frac{1}{3}$$

Recall that the signal power is the square of its rms value. Therefore, the rms value of this signal is $1/\sqrt{3}$.

EXAMPLE 2.2 Determine the power and the rms value of:

(a) $g(t) = C \cos(\omega_0 t + \theta)$

(b) $g(t) = C_1 \cos(\omega_1 t + \theta_1) + C_2 \cos(\omega_2 t + \theta_2) \qquad \omega_1 \neq \omega_2$

(c) $g(t) = De^{j\omega_0 t}$

(a) This is a periodic signal with period $T_0 = 2\pi/\omega_0$. The suitable measure of this signal is its power. Because it is a periodic signal, we may compute its power by averaging its energy over one period $2\pi/\omega_0$. However, for the sake of generality, we shall solve this problem by averaging over an infinitely large time interval using Eq (2.3),

$$\begin{aligned}
P_g &= \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} C^2 \cos^2(\omega_0 t + \theta)\,dt \\
&= \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \frac{C^2}{2}\left[1 + \cos(2\omega_0 t + 2\theta)\right] dt \\
&= \lim_{T\to\infty} \frac{C^2}{2T}\int_{-T/2}^{T/2} dt + \lim_{T\to\infty} \frac{C^2}{2T}\int_{-T/2}^{T/2} \cos(2\omega_0 t + 2\theta)\,dt
\end{aligned}$$

The first term on the right-hand side is equal to $C^2/2$. Moreover, the second term is zero because the integral appearing in this term represents the area under a sinusoid over a very large time interval T with $T \to \infty$. This area is at most equal to the area of half the cycle because of cancellations of the positive and negative areas of a sinusoid. The second term is this area multiplied by $C^2/2T$ with $T \to \infty$. Clearly this term is zero, and

$$P_g = \frac{C^2}{2} \tag{2.6a}$$

This shows that a sinusoid of amplitude C has a power $C^2/2$ regardless of the value of its frequency ω_0 ($\omega_0 \neq 0$) and phase θ. The rms value is $C/\sqrt{2}$. If the signal frequency is zero (dc or a constant signal of amplitude C), the reader can show that the power is C^2.

(b) In this case,

$$\begin{aligned}
P_g &= \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \left[C_1 \cos(\omega_1 t + \theta_1) + C_2 \cos(\omega_2 t + \theta_2)\right]^2 dt \\
&= \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} C_1{}^2 \cos^2(\omega_1 t + \theta_1)\,dt
\end{aligned}$$

$$+\lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} C_2{}^2\cos^2(\omega_2 t+\theta_2)\,dt$$

$$+\lim_{T\to\infty}\frac{2C_1C_2}{T}\int_{-T/2}^{T/2}\cos(\omega_1 t+\theta_1)\cos(\omega_2 t+\theta_2)\,dt$$

Observe that the first and second integrals on the right-hand side are the powers of the two sinusoids, which are $C_1{}^2/2$ and $C_2{}^2/2$ as found in part **(a)**. We now show that the third term on the right-hand side is zero. Using a trigonometric identity, this term, which is the product of the two sinusoids, is equal to the sum of two sinusoids or frequencies $\omega_1+\omega_2$ and $\omega_1-\omega_2$. Thus, the third term is $2C_1C_2/T$ times the sum of the areas under two sinusoids. Now the area under any sinusoid over a large time interval is at most equal to the area under half the cycle because of cancellations of positive and negative areas as argued in part **(a)**. So the third term vanishes because $T\to\infty$, and we have*

$$P_g=\frac{C_1{}^2}{2}+\frac{C_2{}^2}{2} \tag{2.6b}$$

and the rms value is $\sqrt{(C_1{}^2+C_2{}^2)/2}$.

We can readily extend this result to a sum of any number of sinusoids with distinct frequencies ω_n $(\omega_n\neq 0)$. Thus, if

$$g(t)=\sum_{n=1}^{\infty}C_n\cos(\omega_n t+\theta_n)$$

where none of the two sinusoids have identical frequencies, then

$$P_g=\frac{1}{2}\sum_{n=1}^{\infty}C_n{}^2 \tag{2.6c}$$

(c) In this case the signal is complex, and we use Eq. (2.4) to compute the power. However, because this signal is periodic, we need average it only over a period T_0. Thus,

$$P_g=\frac{1}{T_0}\int_0^{T_0}|De^{j\omega_0 t}|^2\,dt$$

Recall that $|e^{j\omega_0 t}|=1$ so that $|De^{j\omega_0 t}|^2=|D|^2$, and

$$P_g=\frac{|D|^2}{T_0}\int_0^{T_0}dt=|D|^2 \tag{2.6d}$$

The rms value is $|D|$.

Comments: In part **(b)** we have shown that the power of the sum of two sinusoids is equal to the sum of the powers of the sinusoids. It appears that the power of $g_1(t)+g_2(t)$ is $P_{g_1}+P_{g_2}$. Be cautioned against such a generalization. All we have proved here is that this is true if the two signals $g_1(t)$ and $g_2(t)$ happen to be sinusoids. It is not true in general. In fact, it is not true even for the sinusoids if the two sinusoids are of the same frequency.

---

* This is true only if $\omega_1\neq\omega_2$. If $\omega_1=\omega_2$, the integrand of the third term is a nonnegative entity, and the integral in the third term $\to\infty$ as $T\to\infty$.

We shall show in Sec. 2.5.3 that only under a certain condition (called orthogonality condition) the power (or energy) of $g_1(t) + g_2(t)$ is equal to the sum of the powers (or energies) of $g_1(t)$ and $g_2(t)$.

2.2 CLASSIFICATION OF SIGNALS

There are several classes of signals. Here we shall consider only the following classes, which are suitable for the scope of this book:

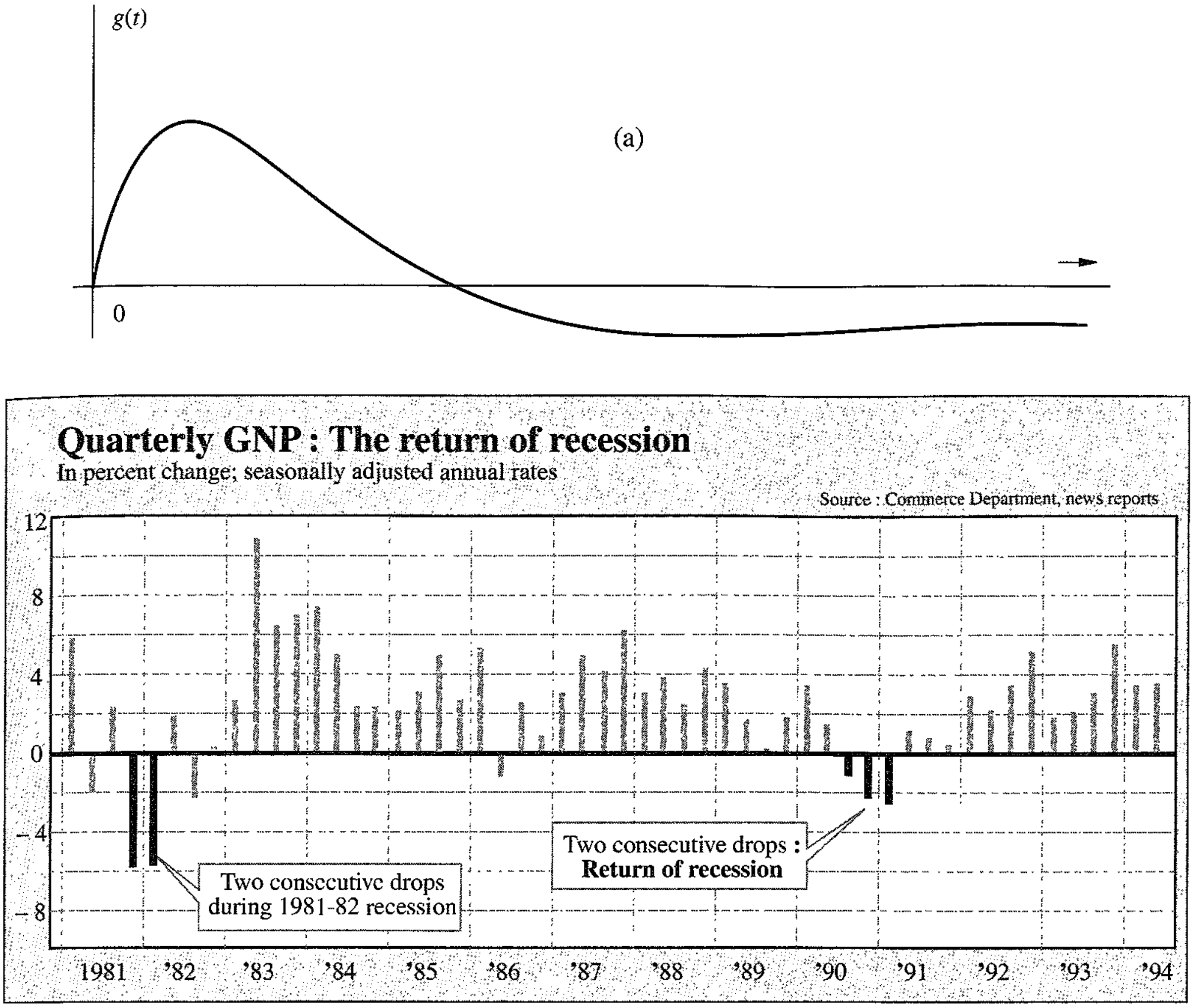
1. Continuous-time and discrete-time signals
2. Analog and digital signals
3. Periodic and aperiodic signals
4. Energy and power signals
5. Deterministic and probabilistic signals

2.2.1 Continuous-Time and Discrete-Time Signals

A signal that is specified for every value of time t (Fig. 2.4a) is a **continuous-time signal**, and a signal that is specified only at discrete values of t (Fig. 2.4b) is a **discrete-time signal**. Telephone and video camera outputs are continuous-time signals, whereas the quarterly gross national product (GNP), monthly sales of a corporation, and stock market daily averages are discrete-time signals.

2.2.2 Analog and Digital Signals

The concept of continuous time is often confused with that of analog. The two are not the same. The same is true of the concepts of discrete time and digital. A signal whose amplitude can take on any value in a continuous range is an **analog signal**. This means that an analog signal amplitude can take on an infinite number of values. A **digital signal**, on the other hand, is one whose amplitude can take on only a finite number of values. Signals associated with a digital computer are digital because they take on only two values (binary signals). For a signal to qualify as digital, the number of values need not be restricted to two. It can be any finite number. A digital signal whose amplitudes can take on M values is an **M-ary** signal of which binary ($M = 2$) is a special case. The terms *continuous time* and *discrete time* qualify the nature of a signal along the time (horizontal) axis. The terms *analog* and *digital*, on the other hand, qualify the nature of the signal amplitude (vertical axis). Figure 2.5 shows examples of various types of signals. It is clear that analog is not necessarily continuous time and digital need not be discrete time. Figure 2.5c shows an example of an analog but discrete-time signal. An analog signal can be converted into a digital signal [analog-to-digital (A/D) conversion] through quantization (rounding off), as explained in Sec. 6.2.



(a)

(b)

Figure 2.4 Continuous-time and discrete-time signals.

2.2.3 Periodic and Aperiodic Signals

A signal $g(t)$ is said to be **periodic** if for some positive constant T_0,

$$g(t) = g(t + T_0) \qquad \text{for all } t \tag{2.7}$$

The *smallest* value of T_0 that satisfies the periodicity condition (2.7) is the **period** of $g(t)$. The signal in Fig. 2.3b is a periodic signal with period 2. A signal is **aperiodic** if it is not periodic. The signal in Fig. 2.3a is aperiodic.

By definition, a periodic signal $g(t)$ remains unchanged when time-shifted by one period. This means that a periodic signal must start at $t = -\infty$ because if it starts at some finite instant, say, $t = 0$, the time-shifted signal $g(t + T_0)$ will start at $t = -T_0$ and $g(t + T_0)$ would not be the same as $g(t)$. Therefore, a *periodic signal, by definition, must start at* $-\infty$ *and continue forever*, as shown in Fig. 2.6. Observe that a periodic signal shifted by an integral multiple of

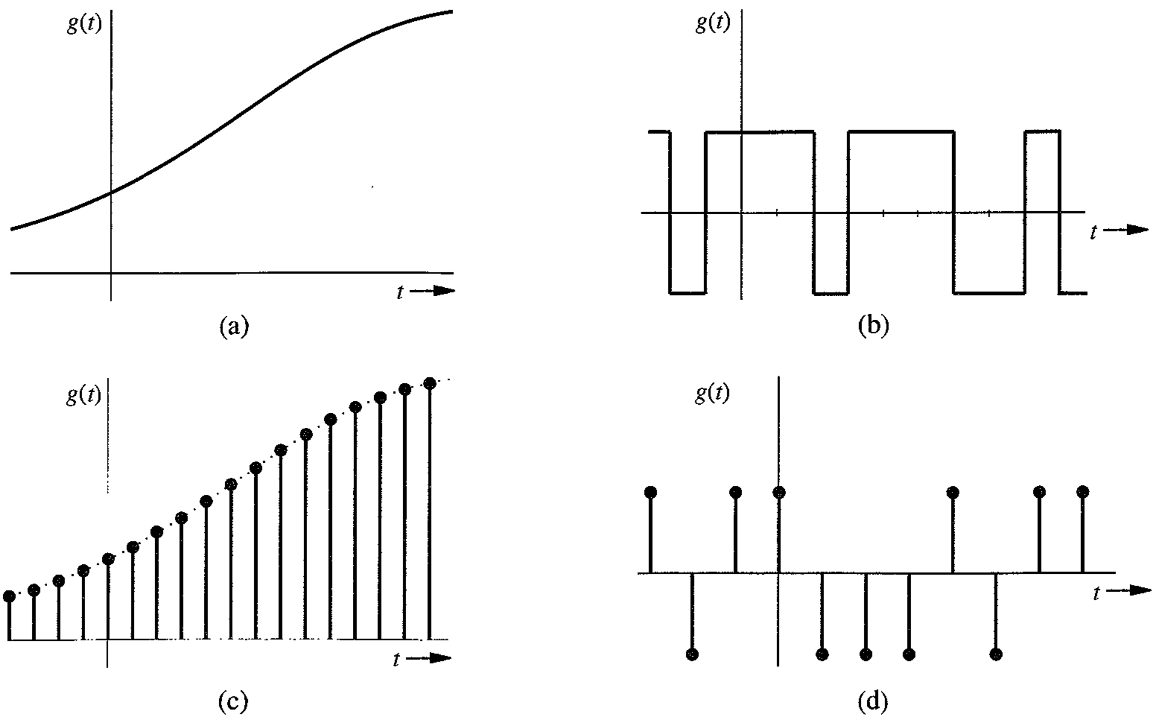


Figure 2.5 Examples of signals. (a) Analog, continuous time. (b) Digital, continuous time. (c) Analog, discrete time. (d) Digital, discrete time.

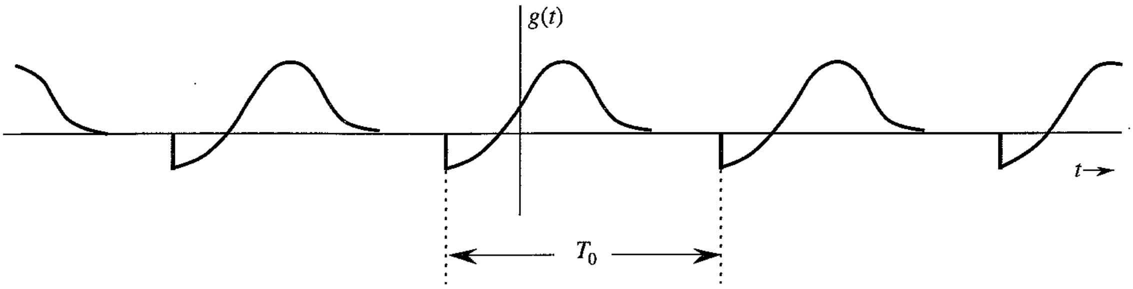


Figure 2.6 Periodic signal of period T_0.

T_0 remains unchanged. Therefore, $g(t)$ may be considered a periodic signal with period mT_0, where m is any integer. However, by definition, the period is the smallest interval that satisfies periodicity condition (2.7). Therefore, T_0 is the period.

The second important property of a periodic signal $g(t)$ is that $g(t)$ can be generated by periodic extension of any segment of $g(t)$ of duration T_0 (the period). This means that we can generate $g(t)$ from any segment of $g(t)$ with a duration of one period by placing this segment and the reproduction thereof end to end ad infinitum on either side. Figure 2.7 shows a periodic signal $g(t)$ of period $T_0 = 6$. The shaded portion of Fig. 2.7a shows a segment of $g(t)$ starting at $t = -1$ and having a duration of one period (6 seconds). This segment, when repeated forever in either direction, results in the periodic signal $g(t)$. Figure 2.7b shows another shaded segment of $g(t)$ of duration T_0 starting at $t = 0$. Again we see that this segment,

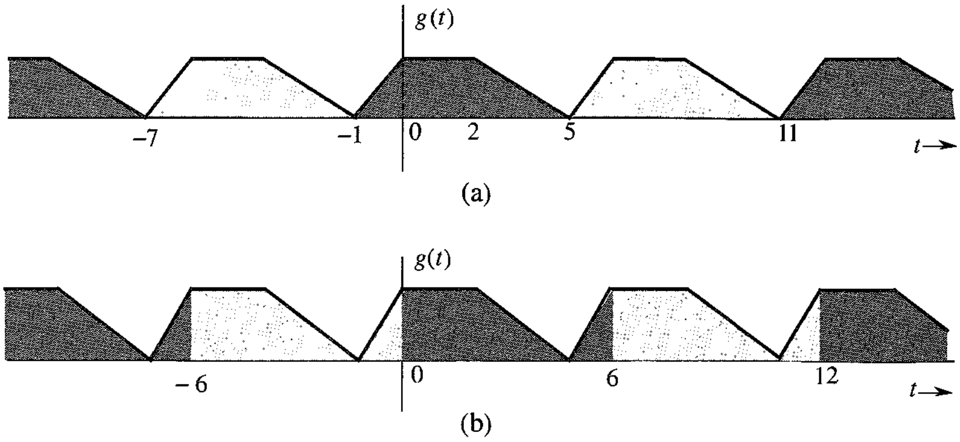


Figure 2.7 Generation of a periodic signal by periodic extension of its segment of one-period duration.

when repeated forever on either side, results in $g(t)$. The reader can verify that this is possible with any segment of $g(t)$ starting at any instant as long as the segment duration is one period.

2.2.4 Energy and Power Signals

A signal with finite energy is an **energy signal**, and a signal with finite power is a **power signal**. In other words, a signal $g(t)$ is an energy signal if

$$\int_{-\infty}^{\infty} |g(t)|^2 \, dt < \infty \tag{2.8}$$

Similarly, a signal with a finite and nonzero power (mean square value) is a power signal. In other words, a signal is a power signal if

$$0 < \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} |g(t)|^2 \, dt < \infty \tag{2.9}$$

Signals in Fig. 2.3a and b are examples of energy and power signals, respectively. Observe that power is the time average of the energy. Since the averaging is over an infinitely large interval, a signal with finite energy has zero power, and a signal with finite power has infinite energy. Therefore, a signal cannot both be an energy and a power signal. If it is one, it cannot be the other. On the other hand, there are signals that are neither energy nor power signals. The ramp signal is such an example.

Comments

Every signal that can be generated in a lab has a finite energy. In other words, *every signal observed in real life is an energy signal. A power signal, on the other hand, must necessarily have an infinite duration.* Otherwise its power, which is its average energy (averaged over an infinitely large interval) will not approach a (nonzero) limit. Obviously it is impossible to generate a true power signal in practice because such a signal has infinite duration and infinite energy.

Also because of periodic repetition, periodic signals for which the area under $|g(t)|^2$ over one period is finite are power signals; however, not all power signals are periodic.

2.2.5 Deterministic and Random Signals

A signal whose physical description is known completely, in either a mathematical form or a graphical form, is a **deterministic signal**. If a signal is known only in terms of probabilistic description, such as mean value, mean squared value, and so on, rather than its complete mathematical or graphical description, is a **random signal**. Most of the noise signals encountered in practice are random signals. All message signals are random signals because, as will be shown later, a signal, to convey information, must have some uncertainty (randomness) about it. The treatment of random signals will be discussed in Chapter 11.

2.3 SOME USEFUL SIGNAL OPERATIONS

We discuss here three useful signal operations: shifting, scaling, and inversion. Since the independent variable in our signal description is time, these operations are discussed as time shifting, time scaling, and time inversion (or folding). However, this discussion is valid for functions having independent variables other than time (e.g., frequency or distance).

2.3.1 Time Shifting

Consider a signal $g(t)$ (Fig. 2.8a) and the same signal delayed by T seconds (Fig. 2.8b), which we shall denote by $\phi(t)$. Whatever happens in $g(t)$ (Fig. 2.8a) at some instant t also happens in $\phi(t)$ (Fig. 2.8b) T seconds later at the instant $t + T$. Therefore,

$$\phi(t+T) = g(t) \tag{2.10}$$

and

$$\phi(t) = g(t-T) \tag{2.11}$$

Therefore, to time-shift a signal by T, we replace t with $t - T$. Thus, $g(t-T)$ represents $g(t)$ time-shifted by T seconds. If T is positive, the shift is to the right (delay). If T is negative, the shift is to the left (advance). Thus, $g(t-2)$ is $g(t)$ delayed (right-shifted) by 2 seconds, and $g(t+2)$ is $g(t)$ advanced (left-shifted) by 2 seconds.

2.3.2 Time Scaling

The compression or expansion of a signal in time is known as **time scaling**. Consider the signal $g(t)$ of Fig. 2.9a. The signal $\phi(t)$ in Fig. 2.9b is $g(t)$ compressed in time by a factor of 2. Therefore, whatever happens in $g(t)$ at some instant t also happens to $\phi(t)$ at the instant $t/2$, so that

$$\phi\left(\frac{t}{2}\right) = g(t) \tag{2.12}$$

and

$$\phi(t) = g(2t) \tag{2.13}$$

Observe that because $g(t) = 0$ at $t = T_1$ and T_2, the same thing must happen in $\phi(t)$ at half

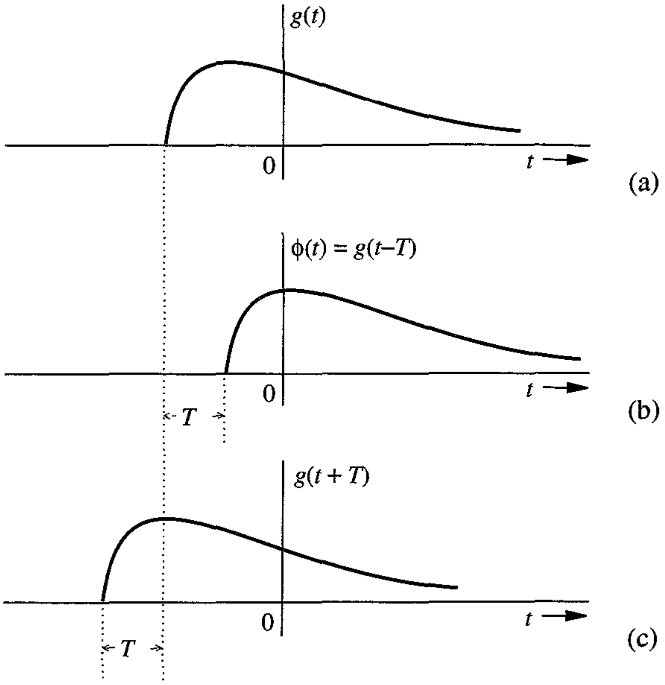

g(t)
0
t
(a)
φ(t) = g(t−T)
0
t
T
(b)
g(t + T)
0
t
T
(c)

Figure 2.8 Time shifting a signal.

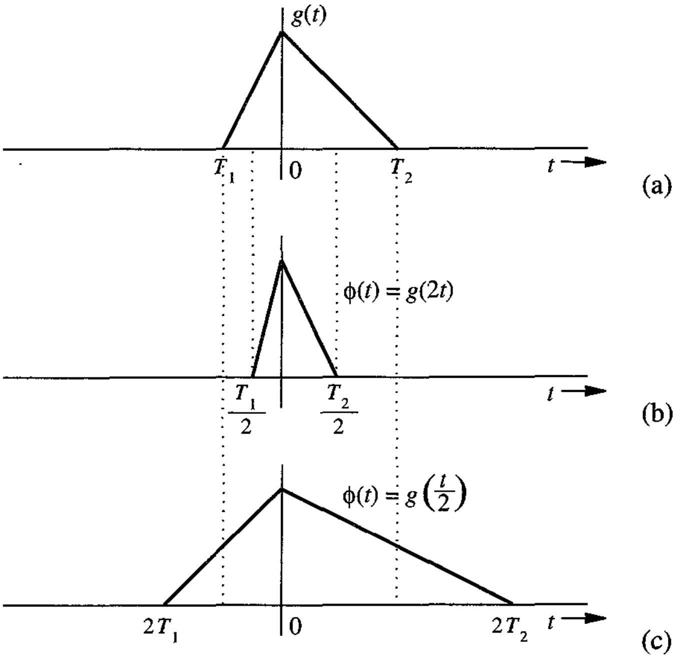

g(t)
T₁
0
T₂
t
(a)
φ(t) = g(2t)
T₁/2
T₂/2
t
(b)
φ(t) = g(t/2)
2T₁
0
2T₂
t
(c)

Figure 2.9 Time scaling a signal.

these values. Therefore, $\phi(t) = 0$ at $t = T_1/2$ and $T_2/2$, as shown in Fig. 2.9b. If $g(t)$ were recorded on a tape and played back at twice the normal recording speed, we would obtain $g(2t)$. In general, if $g(t)$ is compressed in time by a factor a ($a > 1$), the resulting signal $\phi(t)$ is given by

$$\phi(t) = g(at) \tag{2.14}$$

Using a similar argument, we can show that $g(t)$ expanded (slowed down) in time by a factor a ($a > 1$) is given by

$$\phi(t) = g\left(\frac{t}{a}\right) \tag{2.15}$$

Figure 2.9c shows $g(t/2)$, which is $g(t)$ expanded in time by a factor of 2. Note that the signal remains anchored at $t = 0$ during scaling operation (expanding or compressing). In other words, the signal at $t = 0$ remains unchanged. This is because $g(t) = g(at) = g(0)$ at $t = 0$.

In summary, to time-scale a signal by a factor a, we replace t with at. If $a > 1$, the scaling is compression, and if $a < 1$, the scaling is expansion.

EXAMPLE 2.3 Figure 2.10a and b shows the signals $g(t)$ and $z(t)$, respectively. Sketch: **(a)** $g(3t)$; **(b)** $z(t/2)$.

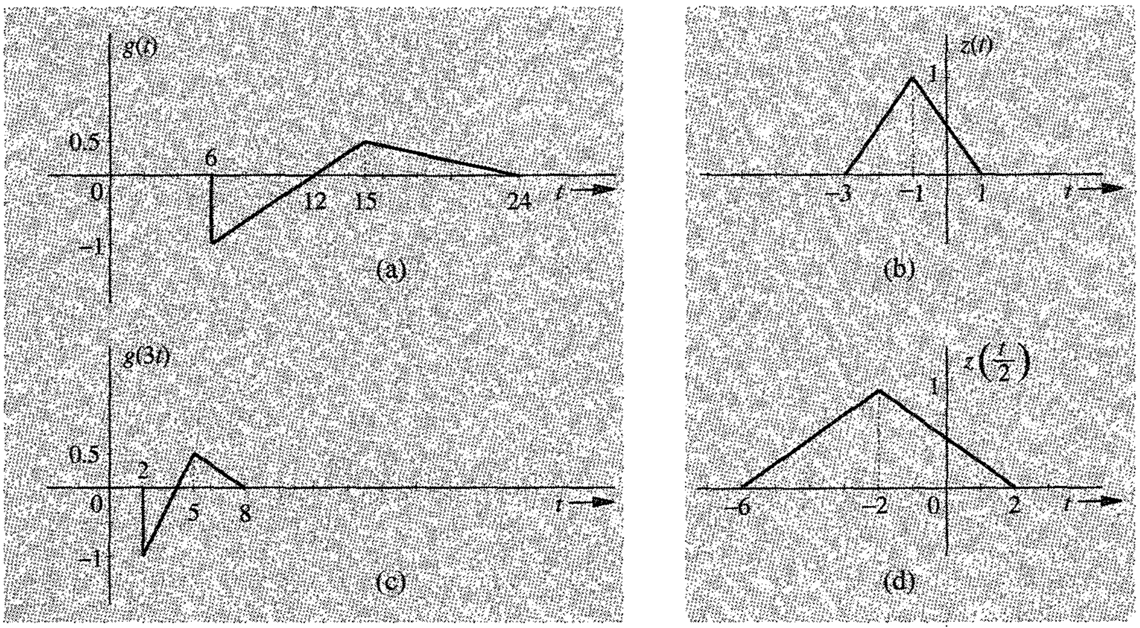


Figure 2.10 Examples of time compression and time expansion of signals.

(a) $g(3t)$ is $g(t)$ compressed by a factor of 3. This means that the values of $g(t)$ at $t = 6$, 12, 15, and 24 occur in $g(3t)$ at the instants $t = 2$, 4, 5, and 8, respectively, as shown in Fig. 2.10c.

(b) $z(t/2)$ is $z(t)$ expanded (slowed down) by a factor of 2. The values of $z(t)$ at $t = 1, -1$, and -3 occur in $z(t/2)$ at instants 2, −2, and −6, respectively, as shown in Fig. 2.10d.

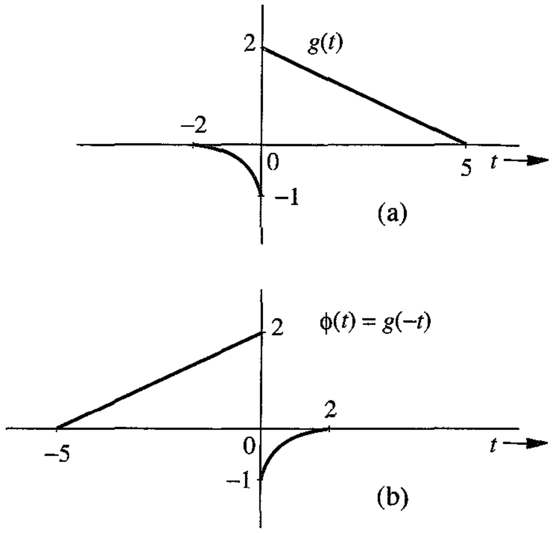


Figure 2.11 Time inversion (reflection) of a signal.

2.3.3 Time Inversion (Time Reversal)

Time inversion may be considered a special case of time scaling with $a = -1$ in Eq. (2.14). Consider the signal $g(t)$ in Fig. 2.11a. We can view $g(t)$ as a rigid wire frame hinged at the vertical axis. To invert $g(t)$, we rotate this frame 180° about the vertical axis. This time inversion or folding [the mirror image of $g(t)$ about the vertical axis] gives us the signal $\phi(t)$ (Fig. 2.11b). Observe that whatever happens in Fig. 2.11a at some instant t also happens in Fig. 2.11b at the instant $-t$. Therefore,

$$\phi(-t) = g(t)$$

and

$$\phi(t) = g(-t) \tag{2.16}$$

Therefore, to time-invert a signal we replace t with $-t$. Thus, the time inversion of signal $g(t)$ yields $g(-t)$. Consequently, the mirror image of $g(t)$ about the vertical axis is $g(-t)$. Recall also that the mirror image of $g(t)$ about the horizontal axis is $-g(t)$.

EXAMPLE 2.4 For the signal $g(t)$ shown in Fig. 2.12a, sketch $g(-t)$.

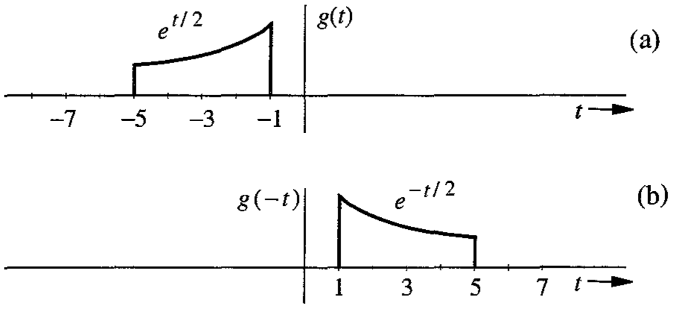


Figure 2.12 Example of time inversion.

The instants −1 and −5 in $g(t)$ are mapped into instants 1 and 5 in $g(-t)$. If $g(t) = e^{t/2}$, then $g(-t) = e^{-t/2}$. The signal $g(-t)$ is shown in Fig. 2.12b.

2.4 UNIT IMPULSE FUNCTION

The unit impulse function $\delta(t)$ is one of the most important functions in the study of signals and systems. This function was first defined by P. A. M. Dirac as

$$\delta(t) = 0 \qquad t \neq 0$$
$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1 \tag{2.17}$$

We can visualize an impulse as a tall, narrow rectangular pulse of unit area, as shown in Fig. 2.13b. The width of this rectangular pulse is some very small value ϵ. Its height is a very large value $1/\epsilon$ in the limit as $\epsilon \to 0$. The unit impulse therefore can be regarded as a rectangular pulse with a width that has become infinitesimally small, a height that has become infinitely large, and an overall area that has been maintained at unity.* Thus, $\delta(t) = 0$ everywhere except at $t = 0$, where it is undefined. For this reason a unit impulse is represented by the spearlike symbol in Fig. 2.13a.

Multiplication of a Function by an Impulse

Let us now consider what happens when we multiply the unit impulse $\delta(t)$ by a function $\phi(t)$ that is known to be continuous at $t = 0$. Since the impulse exists only at $t = 0$, and the value of $\phi(t)$ at $t = 0$ is $\phi(0)$, we obtain

$$\phi(t)\delta(t) = \phi(0)\delta(t) \tag{2.18a}$$

Similarly, if $\phi(t)$ is multiplied by an impulse $\delta(t - T)$ (an impulse located at $t = T$), then

$$\phi(t)\delta(t - T) = \phi(T)\delta(t - T) \tag{2.18b}$$

provided $\phi(t)$ is continuous at $t = T$.

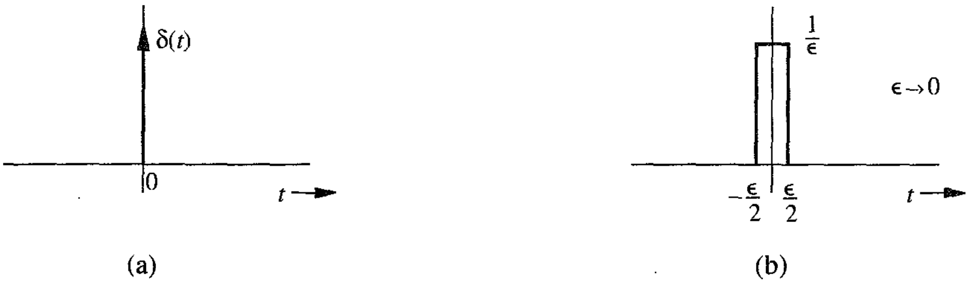


Figure 2.13 Unit impulse and its approximation.

---

* The impulse function can also be approximated by other pulses, such as an exponential pulse, a triangular pulse, or a Gaussian pulse.

Sampling Property of the Unit Impulse Function

From Eq. (2.18a) it follows that

$$\int_{-\infty}^{\infty} \phi(t)\delta(t)\, dt = \phi(0) \int_{-\infty}^{\infty} \delta(t)\, dt$$
$$= \phi(0) \tag{2.19a}$$

provided $\phi(t)$ is continuous at $t = 0$. This result means that *the area under the product of a function with an impulse* $\delta(t)$ *is equal to the value of that function at the instant where the unit impulse is located.* This property is very important and useful, and is known as the **sampling**, or **sifting, property** of the unit impulse.

From Eq. (2.18b) it follows that

$$\int_{-\infty}^{\infty} \phi(t)\delta(t - T)\, dt = \phi(T) \tag{2.19b}$$

Equation (2.19b) is just another form of sampling or sifting property. In the case of Eq. (2.19b), the impulse $\delta(t - T)$ is located at $t = T$. Therefore, the area under $\phi(t)\delta(t - T)$ is $\phi(T)$, the value of $\phi(t)$ at the instant where the impulse is located (at $t = T$). In these derivations we have assumed that the function is continuous at the instant where the impulse is located.

Unit Impulse as a Generalized Function

The definition of the unit impulse function [Eq. (2.17)] leads to a nonunique function.[1] Moreover, $\delta(t)$ is not even a true function in the ordinary sense. An ordinary function is specified by its values for all time t. The impulse function is zero everywhere except at $t = 0$, and at this only interesting part of its range it is undefined. In a more rigorous approach, the impulse function is defined not as an ordinary function but as a **generalized function**, where $\delta(t)$ is defined by Eqs. (2.19). We say nothing about what the impulse function is or what it looks like. Instead, it is defined in terms of the effect it has on a test function $\phi(t)$. We define a unit impulse as a function for which the area under its product with a function $\phi(t)$ is equal to the value of the function $\phi(t)$ at the instant where the impulse is located. Recall that the sampling property [Eqs. (2.19)] is the consequence of the classical (Dirac) definition of impulse in Eq. (2.17). In contrast, *the sampling property [Eqs. (2.19)] defines the impulse function in the generalized function approach.*

Unit Step Function $u(t)$

Another familiar and useful function is the **unit step function** $u(t)$, defined by (Fig. 2.14a)

$$u(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases}$$

If we want a signal to start at $t = 0$ (so that it has a value of zero for $t < 0$), we only need to multiply the signal with $u(t)$. A signal that does not start before $t = 0$ is called a **causal signal**. In other words, $g(t)$ is a causal signal if

$$g(t) = 0 \qquad t < 0$$

The signal e^{-at} represents an exponential that starts at $t = -\infty$. If we want this signal to start at $t = 0$ (the causal form), it can be described as $e^{-at}u(t)$ (Fig. 2.14b). From Fig. 2.13b, we

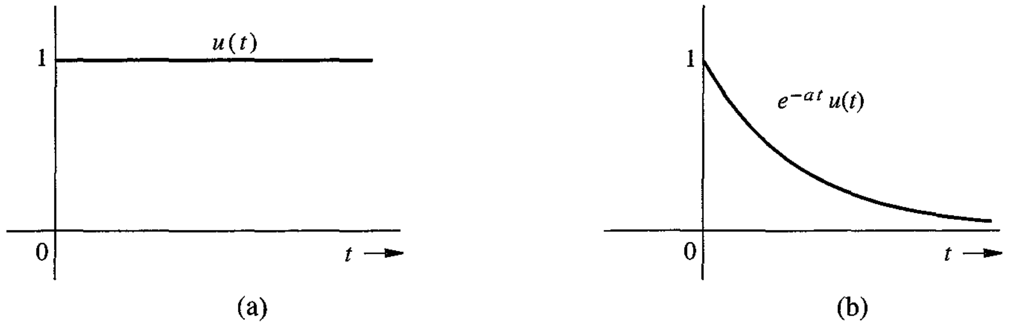


Figure 2.14 (a) Unit step function $u(t)$. (b) Causal exponential $e^{-at}u(t)$.

observe that the area from $-\infty$ to t under the limiting form of $\delta(t)$ is zero if $t < 0$ and unity if $t \geq 0$. Consequently,

$$\int_{-\infty}^{t} \delta(\tau)\, d\tau = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases}$$
$$= u(t) \tag{2.20a}$$

From this result it follows that

$$\frac{du}{dt} = \delta(t) \tag{2.20b}$$

2.5 SIGNALS AND VECTORS

There is a perfect analogy between signals and vectors. The analogy is so strong that the term "analogy" understates the reality. Signals are not just *like* vectors. Signals *are* vectors. A vector can be represented as a sum of its components in a variety of ways, depending on the choice of coordinate system. A signal can also be represented as a sum of its components in a variety of ways. Let us begin with some basic vector concepts and then apply those concepts to signals.

2.5.1 Component of a Vector

A vector is specified by its magnitude and its direction. We shall denote all vectors by boldface type. For example, $\mathbf{x}$ is a certain vector with magnitude or length $|\mathbf{x}|$. Consider two vectors $\mathbf{g}$ and $\mathbf{x}$, as shown in Fig. 2.15. Let the component of $\mathbf{g}$ along $\mathbf{x}$ be $c\mathbf{x}$. Geometrically the component of $\mathbf{g}$ along $\mathbf{x}$ is the projection of $\mathbf{g}$ on $\mathbf{x}$, and is obtained by drawing a perpendicular from the tip of $\mathbf{g}$ on the vector $\mathbf{x}$, as shown in Fig. 2.15. What is the mathematical significance of a component of a vector along another vector? As seen from Fig. 2.15, the vector $\mathbf{g}$ can be expressed in terms of vector $\mathbf{x}$ as

$$\mathbf{g} = c\mathbf{x} + \mathbf{e} \tag{2.21}$$

However, this is not the only way to express $\mathbf{g}$ in terms of $\mathbf{x}$. Figure 2.16 shows two of the infinite other possibilities. From Fig. 2.16a and b, we have

$$\mathbf{g} = c_1\mathbf{x} + \mathbf{e}_1 = c_2\mathbf{x} + \mathbf{e}_2 \tag{2.22}$$

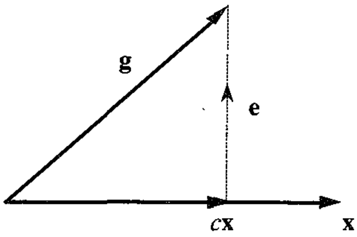


Figure 2.15 Component (projection) of a vector along another vector.

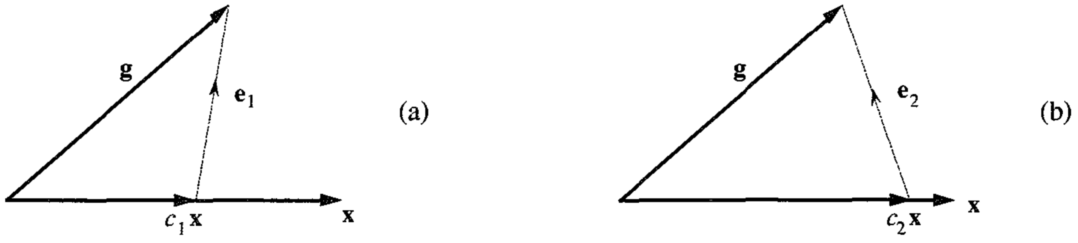


Figure 2.16 Approximation of a vector in terms of another vector.

In each of these three representations (Figs. 2.15 and 2.16) $\mathbf{g}$ is represented in terms of $\mathbf{x}$ plus another vector, called the **error vector**. If we approximate $\mathbf{g}$ by $c\mathbf{x}$ (Fig. 2.15),

$$\mathbf{g} \simeq c\mathbf{x} \tag{2.23}$$

the error in this approximation is the vector $\mathbf{e} = \mathbf{g} - c\mathbf{x}$. Similarly, the errors in the approximations in Fig. 2.16a and b are $\mathbf{e}_1$ and $\mathbf{e}_2$. What is unique about the approximation in Fig. 2.15 is that the error vector is the smallest. We can now define mathematically the component of a vector $\mathbf{g}$ along vector $\mathbf{x}$ to be $c\mathbf{x}$, where c is chosen to minimize the length of the error vector $\mathbf{e} = \mathbf{g} - c\mathbf{x}$. For convenience we define the dot (inner or scalar) product of two vectors $\mathbf{g}$ and $\mathbf{x}$ as

$$\mathbf{g} \cdot \mathbf{x} = |\mathbf{g}||\mathbf{x}| \cos \theta \tag{2.24}$$

where θ is the angle between vectors $\mathbf{g}$ and $\mathbf{x}$. Using this definition, we can express $|\mathbf{x}|$, the length of a vector $\mathbf{x}$, as

$$|\mathbf{x}|^2 = \mathbf{x} \cdot \mathbf{x} \tag{2.25}$$

Now, the length of the component of $\mathbf{g}$ along $\mathbf{x}$ is $|\mathbf{g}| \cos \theta$, but it is also $c|\mathbf{x}|$. Therefore,

$$c|\mathbf{x}| = |\mathbf{g}| \cos \theta$$

Multiplying both sides by $|\mathbf{x}|$ yields

$$c|\mathbf{x}|^2 = |\mathbf{g}||\mathbf{x}| \cos \theta = \mathbf{g} \cdot \mathbf{x}$$

and

$$c = \frac{\mathbf{g} \cdot \mathbf{x}}{\mathbf{x} \cdot \mathbf{x}} = \frac{1}{|\mathbf{x}|^2}\, \mathbf{g} \cdot \mathbf{x} \tag{2.26}$$

From Fig. 2.15, it is apparent that when **g** and **x** are perpendicular, or orthogonal, then **g** has a zero component along **x**; consequently, $c = 0$. Keeping an eye on Eq. (2.26), we therefore define **g** and **x** to be **orthogonal** if the inner (scalar or dot) product of the two vectors is zero, that is, if

$$\mathbf{g} \cdot \mathbf{x} = 0 \tag{2.27}$$

2.5.2 Component of a Signal

The concepts of vector component and orthogonality can be extended to signals. Consider the problem of approximating a real signal $g(t)$ in terms of another real signal $x(t)$ over an interval $[t_1, t_2]$:

$$g(t) \simeq cx(t) \qquad t_1 \leq t \leq t_2 \tag{2.28}$$

The error $e(t)$ in this approximation is

$$e(t) = \begin{cases} g(t) - cx(t) & t_1 \leq t \leq t_2 \\ 0 & \text{otherwise} \end{cases} \tag{2.29}$$

We now select some criterion for the "best approximation." We know that the signal energy is one possible measure of a signal size. For best approximation, we need to minimize the error signal, that is, minimize its size, which is its energy E_e over the interval $[t_1, t_2]$, given by

$$\begin{aligned} E_e &= \int_{t_1}^{t_2} e^2(t)\, dt \\ &= \int_{t_1}^{t_2} [g(t) - cx(t)]^2\, dt \end{aligned}$$

Note that the right-hand side is a definite integral with t as the dummy variable. Hence, E_e is a function of the parameter c (not t) and E_e is minimum for some choice of c. To minimize E_e, a necessary condition is

$$\frac{dE_e}{dc} = 0 \tag{2.30}$$

or

$$\frac{d}{dc}\left[\int_{t_1}^{t_2} [g(t) - cx(t)]^2\, dt\right] = 0$$

Expanding the squared term inside the integral, we obtain

$$\frac{d}{dc}\left[\int_{t_1}^{t_2} g^2(t)\, dt\right] - \frac{d}{dc}\left[2c\int_{t_1}^{t_2} g(t)x(t)\, dt\right] + \frac{d}{dc}\left[c^2\int_{t_1}^{t_2} x^2(t)\, dt\right] = 0$$

from which we obtain

$$-2\int_{t_1}^{t_2} g(t)x(t)\, dt + 2c\int_{t_1}^{t_2} x^2(t)\, dt = 0$$

and

$$c = \frac{\int_{t_1}^{t_2} g(t)x(t)\,dt}{\int_{t_1}^{t_2} x^2(t)\,dt} = \frac{1}{E_x}\int_{t_1}^{t_2} g(t)x(t)\,dt \tag{2.31}$$

We observe a remarkable similarity between the behavior of vectors and signals, as indicated by Eqs. (2.26) and (2.31). It is evident from these two parallel expressions that *the area under the product of two signals corresponds to the inner (scalar or dot) product of two vectors.* In fact, the area under the product of $g(t)$ and $x(t)$ is called the **inner product** of $g(t)$ and $x(t)$, and is denoted by (f, g). The energy of a signal is the inner product of a signal with itself, and corresponds to the vector length squared (which is the inner product of the vector with itself).

To summarize our discussion, if a signal $g(t)$ is approximated by another signal $x(t)$ as

$$g(t) \simeq cx(t)$$

then the optimum value of c that minimizes the energy of the error signal in this approximation is given by Eq. (2.31).

Taking our clue from vectors, we say that a signal $g(t)$ contains a component $cx(t)$, where c is given by Eq. (2.31). Note that in vector terminology, $cx(t)$ is the projection of $g(t)$ on $x(t)$. Continuing with the analogy, we say that if the component of a signal $g(t)$ of the form $x(t)$ is zero (that is, $c = 0$), the signals $g(t)$ and $x(t)$ are orthogonal over the interval $[t_1, t_2]$. Therefore, we define the real signals $g(t)$ and $x(t)$ to be orthogonal over the interval $[t_1, t_2]$ if*

$$\int_{t_1}^{t_2} g(t)x(t)\,dt = 0 \tag{2.32}$$

EXAMPLE 2.5 For the square signal $g(t)$ shown in Fig. 2.17 find the component in $g(t)$ of the form $\sin t$. In other words, approximate $g(t)$ in terms of $\sin t$:

$$g(t) \simeq c \sin t \qquad 0 \le t \le 2\pi$$

so that the energy of the error signal is minimum.

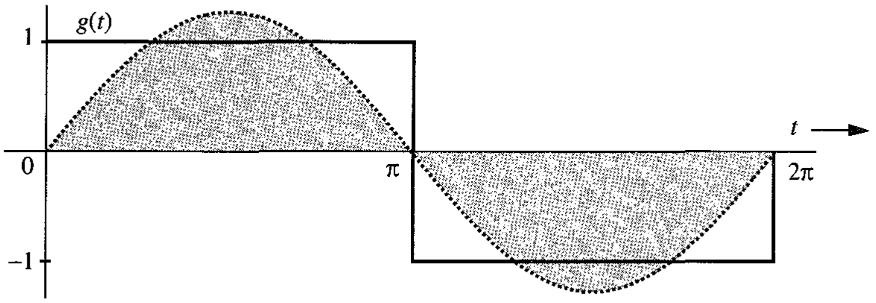


Figure 2.17 Approximation of a square signal in terms of a single sinusoid.

---

[*] For complex signals the definition is modified as in Eq. (2.40).

In this case,

$$x(t) = \sin\, t \qquad \text{and} \qquad E_x = \int_0^{2\pi} \sin^2 t\, dt = \pi$$

From Eq. (2.31), we find

$$c = \frac{1}{\pi}\int_0^{2\pi} g(t)\sin\, t\, dt = \frac{1}{\pi}\left[\int_0^{\pi} \sin\, t\, dt + \int_{\pi}^{2\pi} -\sin\, t\, dt\right] = \frac{4}{\pi} \tag{2.33}$$

Therefore,

$$g(t) \simeq \frac{4}{\pi}\sin\, t \tag{2.34}$$

represents the best approximation of $g(t)$ by the function sin t, which will minimize the error energy. This sinusoidal component of $g(t)$ is shown shaded in Fig. 2.17. By analogy with vectors, we say that the square function $g(t)$ shown in Fig. 2.17 has a component of signal sin t and that the magnitude of this component is $4/\pi$.

2.5.3 Orthogonality in Complex Signals

So far we have restricted ourselves to real functions of t. To generalize the results to complex functions of t, consider again the problem of approximating a function $g(t)$ by a function $x(t)$ over an interval $(t_1 \leq t \leq t_2)$:

$$g(t) \simeq cx(t) \tag{2.35}$$

where $g(t)$ and $x(t)$ now can be complex functions of t. Recall that the energy E_x of the complex signal $x(t)$ over an interval $[t_1,\ t_2]$ is

$$E_x = \int_{t_1}^{t_2} |x(t)|^2\, dt$$

In this case, both the coefficient c and the error

$$e(t) = g(t) - cx(t) \tag{2.36}$$

are complex (in general). For the best approximation, we choose c such that we minimize E_e, the energy of the error signal $e(t)$, given by

$$E_e = \int_{t_1}^{t_2} |g(t) - cx(t)|^2\, dt \tag{2.37}$$

Recall also that

$$|u + v|^2 = (u + v)(u^* + v^*) = |u|^2 + |v|^2 + u^*v + uv^* \tag{2.38}$$

Using this result, we can, after some manipulation, express the integral E_e in Eq. (2.37) as

$$E_e = \int_{t_1}^{t_2} |g(t)|^2\, dt - \left|\frac{1}{\sqrt{E_x}}\int_{t_1}^{t_2} g(t)x^*(t)\, dt\right|^2 + \left|c\sqrt{E_x} - \frac{1}{\sqrt{E_x}}\int_{t_1}^{t_2} g(t)x^*(t)\, dt\right|^2$$

Since the first two terms on the right-hand side are independent of c, it is clear that E_e is minimized by choosing c such that the third term is zero. This yields

$$c = \frac{1}{E_x} \int_{t_1}^{t_2} g(t)x^*(t)\, dt \tag{2.39}$$

In light of this result, we need to redefine orthogonality for the complex case as follows: Two complex functions $x_1(t)$ and $x_2(t)$ are orthogonal over an interval $(t \le t_1 \le t_2)$ if

$$\int_{t_1}^{t_2} x_1(t)x_2^*(t)\, dt = 0 \qquad \text{or} \qquad \int_{t_1}^{t_2} x_1^*(t)x_2(t)\, dt = 0 \tag{2.40}$$

Either equality suffices. This is a general definition of orthogonality, which reduces to Eq. (2.32) when the functions are real.

Energy of the Sum of Orthogonal Signals

We know that the length of the sum of two orthogonal vectors is equal to the sum of the lengths squared of the two vectors. Thus, if vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, and if $\mathbf{z} = \mathbf{x} + \mathbf{y}$, then

$$|\mathbf{z}|^2 = |\mathbf{x}|^2 + |\mathbf{y}|^2$$

We have a similar result for signals. The energy of the sum of two orthogonal signals is equal to the sum of the energies of the two signals. Thus, if signals $x(t)$ and $y(t)$ are orthogonal over an interval $[t_1, t_2]$, and if $z(t) = x(t) + y(t)$, then

$$E_z = E_x + E_y \tag{2.41}$$

We now prove this result for complex signals of which real signals are a special case. From Eq. (2.38) it follows that

$$\begin{aligned}\int_{t_1}^{t_2} |x(t) + y(t)|^2\, dt &= \int_{t_1}^{t_2} |x(t)|^2\, dt + \int_{t_1}^{t_2} |y(t)|^2\, dt + \int_{t_1}^{t_2} x(t)y^*(t)\, dt \int_{t_1}^{t_2} x^*(t)y(t)\, dt \\ &= \int_{t_1}^{t_2} |x(t)|^2\, dt + \int_{t_1}^{t_2} |y(t)|^2\, dt \end{aligned} \tag{2.42}$$

The last result follows from the fact that because of orthogonality, the two integrals of the cross products $x(t)y^*(t)$ and $x^*(t)y(t)$ are zero. This result can be extended to the sum of any number of mutually orthogonal signals.

2.6 SIGNAL COMPARISON: CORRELATION

Section 2.5 has prepared the foundation for signal comparison. Here again, we can benefit by considering the concept of vector comparison. Two vectors $\mathbf{g}$ and $\mathbf{x}$ are similar if $\mathbf{g}$ has a large component along $\mathbf{x}$. In other words, if c in Eq. (2.26) is large, the vectors $\mathbf{g}$ and $\mathbf{x}$ are similar. We could consider c as a quantitative measure of similarity between $\mathbf{g}$ and $\mathbf{x}$. Such a measure, however, would be defective. The amount of similarity between $\mathbf{g}$ and $\mathbf{x}$ should be independent of the lengths of $\mathbf{g}$ and $\mathbf{x}$. If we double the length of $\mathbf{g}$, for example, the amount of similarity between **g** and **x** should not change. From Eq. (2.26), however, we see that doubling **g** doubles the value of c (whereas doubling **x** halves the value of c). Our measure is clearly faulty. Similarity between two vectors is indicated by the angle θ between the vectors. The smaller the θ, the larger is the similarity, and vice versa. The amount of similarity can therefore be conveniently measured by cos θ. The larger the cos θ, larger is the similarity between the two vectors. Thus, a suitable measure would be $c_n = \cos\theta$, which is given by

$$c_n = \cos\theta = \frac{\mathbf{g}\cdot\mathbf{x}}{|\mathbf{g}|\,|\mathbf{x}|} \tag{2.43}$$

We can readily verify that this measure is independent of the lengths of **g** and **x**. This similarity measure c_n is known as the **correlation coefficient**. Observe that

$$-1 \le c_n \le 1 \tag{2.44}$$

Thus, the magnitude of c_n is never greater than unity. If the two vectors are aligned, the similarity is maximum ($c_n = 1$). Two vectors aligned in opposite directions have maximum dissimilarity ($c_n = -1$). If the two vectors are orthogonal, the similarity is zero.

We use the same argument in defining a similarity index (the correlation coefficient) for signals. We shall consider the signals over the entire time interval from $-\infty$ to ∞. To make c in Eq. (2.31) independent of the energies (sizes) of $g(t)$ and $x(t)$, we must normalize c by normalizing the two signals to have unit energies. Thus, the appropriate similarity index c_n analogous to Eq. (2.43) is given by

$$c_n = \frac{1}{\sqrt{E_g E_x}} \int_{-\infty}^{\infty} g(t)x(t)\,dt \tag{2.45}$$

Observe that multiplying either $g(t)$ or $x(t)$ by any constant has no effect on this index. It is independent of the sizes (energies) of $g(t)$ and $x(t)$. Using Schwarz's inequality (proved in Appendix B),[*] we can show that the magnitude of c_n is never greater than 1:

$$-1 \le c_n \le 1 \tag{2.46}$$

Best Friends, Worst Enemies, and Complete Strangers

We can readily verify that if $g(t) = Kx(t)$, then $c_n = 1$ when K is any positive constant, and $c_n = -1$ when K is any negative constant. Also $c_n = 0$ if $g(t)$ and $x(t)$ are orthogonal. Thus, the maximum similarity [when $g(t) = Kx(t)$] is indicated by $c_n = 1$, the maximum dissimilarity [when $g(t) = -Kx(t)$] is indicated by $c_n = -1$. When the two signals are orthogonal, the similarity is zero. Qualitatively speaking, we may view orthogonal signals as unrelated signals. Note that maximum dissimilarity is different from unrelatedness qualitatively. For example, we have the best friends ($c_n = 1$), the worst enemies ($c_n = -1$), and complete strangers, who do not care whether we exist or not ($c_n = 0$). The worst enemies are not strangers but, in many ways, people who think like us, only in opposite ways.

---

[*] The Schwarz inequality states that for two real energy signals $g(t)$ and $x(t)$,

$$\left[\int_{-\infty}^{\infty} g(t)x(t)\,dt\right]^2 \le E_g E_x \tag{2.45n}$$

with equality if and only if $x(t) = Kg(t)$, where K is an arbitrary constant. There is also a similar inequality for complex signals.

We can readily extend this discussion to complex signal comparison. We generalize the definition of c_n to include complex signals as

$$c_n = \frac{1}{\sqrt{E_g E_x}} \int_{-\infty}^{\infty} g(t)x^*(t)\, dt \tag{2.47}$$

EXAMPLE 2.6 Find the correlation coefficient c_n between the pulse $x(t)$ and the pulses $g_i(t)$, $i = 1, 2, 3, 4, 5,$ and 6, shown in Fig. 2.18.

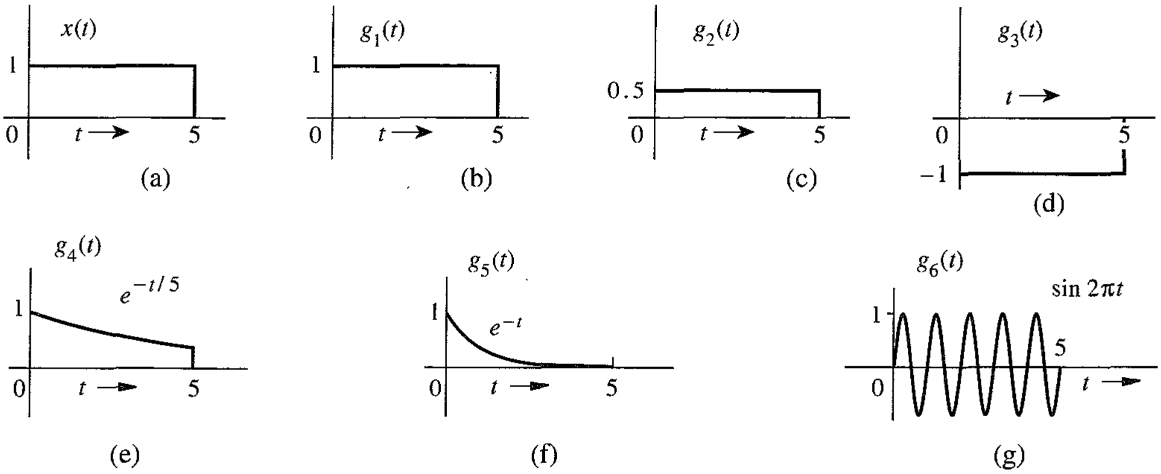


Figure 2.18 Signals for Example 2.6.

We shall compute c_n using Eq. (2.45) for each of the six cases. Let us first compute the energies of all the signals,

$$E_x = \int_0^5 x^2(t)\, dt = \int_0^5 dt = 5$$

In the same way we find $E_{g_1} = 5$, $E_{g_2} = 1.25$, and $E_{g_3} = 5$. Also to determine E_{g_4} and E_{g_5}, we determine the energy E of $e^{-at}u(t)$ over the interval $t = 0$ to T:

$$E = \int_0^T \left(e^{-at}\right)^2 dt = \int_0^T e^{-2at}\, dt = \frac{1}{2a}(1 - e^{-2aT})$$

For $g_4(t)$, $a = 1/5$ and $T = 5$. Therefore, $E_{g_4} = 2.1617$. For $g_5(t)$, $a = 1$ and $T = \infty$. Therefore, $E_{g_5} = 0.5$. The energy of E_{g_6} is given by

$$E_{g_6} = \int_0^5 \sin^2 2\pi t\, dt = 2.5$$

Using Eq. (2.25), the correlation coefficients for the six cases are found as

(1) $\frac{1}{\sqrt{5\,5}} \int_0^5 dt = 1$ **(2)** $\frac{1}{\sqrt{1.25\,5}} \int_0^5 (0.5)\, dt = 1$

(3) $\frac{1}{\sqrt{5\,5}} \int_0^5 (-1)\, dt = -1$ **(4)** $\frac{1}{\sqrt{2.1617\,5}} \int_0^5 e^{-t/5}\, dt = 0.961$

(5) $\frac{1}{\sqrt{0.5\,5}} \int_0^5 e^{-t}\, dt = 0.628$ **(6)** $\frac{1}{\sqrt{2.5\,5}} \int_0^5 \sin 2\pi t\, dt = 0$

Comments: Because $g_1(t) = x(t)$, the two signals have the maximum possible similarity, and $c_n = 1$. However, the signal $g_2(t)$ also shows maximum possible similarity with $c_n = 1$. This is because we have defined c_n to measure the similarity of the wave shapes, and it is independent of the amplitude (strength) of the signals compared. The signal $g_2(t)$ is identical to $x(t)$ in shape; only the amplitude (strength) is different. Hence, $c_n = 1$. The signal $g_3(t)$, on the other hand, has the maximum possible dissimilarity with $x(t)$ because it is equal to $-x(t)$. For $g_4(t)$, $c_n = 0.961$, implying a high degree of similarity with $x(t)$. This is reasonable because $g_4(t)$ is very similar to $x(t)$ over the duration of $x(t)$ (for $0 \leq t \leq 5$). Just by inspection, we notice that the variations or changes in both $x(t)$ and $g_4(t)$ are at similar rates. Such is not the case with $g_5(t)$, where we notice that variations in $g_5(t)$ are generally at a higher rate than those in $x(t)$. There is still considerable similarity: Both signals always remain positive and show no oscillations. Both signals have zero or negligible strength beyond $t = 5$. Thus, $g_5(t)$ is similar to $x(t)$, but not as similar as $g_4(t)$. This is why $c_n = 0.628$ for $g_5(t)$. The signal $g_6(t)$ is orthogonal to $x(t)$, so that $c_n = 0$. This appears to indicate that the dissimilarity in this case is not as strong as that of $g_3(t)$, for which $c_n = -1$. This may seem odd because $g_3(t)$ appears more similar to $x(t)$ than does $g_6(t)$. The dissimilarity between $x(t)$ and $g_3(t)$ is of the nature of antipathy (the worst enemy); in a way they are very similar, but in opposite ways. On the other hand, the dissimilarity of $x(t)$ with $g_6(t)$ stems from the fact that they are almost of different species or from different planets; it is of the nature of being strangers to each other. Hence, the dissimilarity of $x(t)$ with $g_6(t)$ rates lower than that with $g_3(t)$.

2.6.1 Application to Signal Detection

Correlation between two signals is an extremely important concept, which measures the degree of similarity (agreement or alignment) between the two signals. This concept is widely used for signal processing applications in radar, sonar, digital communication, electronic warfare, and many others.

We explain this concept by an example of radar where a signal pulse is transmitted in order to detect a suspected target. If a target is present, the pulse will be reflected by it. If a target is not present, there will be no reflected pulse, just a noise. By detecting the presence or absence of the reflected pulse we confirm the presence or absence of a target. The crucial problem in this procedure is to detect the heavily attenuated, reflected pulse (of known waveform) buried in the unwanted noise signal. Correlation of the received pulse with the transmitted pulse can be of great help in this situation. A similar situation exists in digital communication, where we are required to detect the presence of one of the two known waveforms in the presence of noise.

We now explain qualitatively how signal detection using the correlation technique is accomplished. Consider the case of binary communication, where two known waveforms are received in a random sequence. Each time we receive a pulse, our task is to determine which of the two (known) waveforms is received. To make the detection easier, we must make the two pulses as dissimilar as possible, which means that we should select one pulse as the negative of the other pulse. This gives the highest dissimilarity ($c_n = -1$). This scheme is sometimes called the **antipodal** scheme. We can also use orthogonal pulses, which result in $c_n = 0$. In practice both these options are used, although the antipodal one is best in terms of distinguishability between the two pulses.

Let us consider the antipodal scheme in which the two pulses are $p(t)$ and $-p(t)$. The correlation coefficient c_n of these pulses is -1. Assume that there is no noise nor any other imperfections in the transmission. The receiver consists of a correlator that computes the correlation between $p(t)$ and the received pulse. If the correlation is 1, we decide that $p(t)$ is received; if the correlation is -1, we decide that $-p(t)$ is received. Because of the maximum possible dissimilarity between the two pulses, detection is easier. In practice, however, there are several imperfections. There is always an unwanted signal (noise) superimposed on the received pulses. Moreover, during transmission, pulses get distorted and dispersed (spread out) in time. Consequently, the correlation coefficient is no longer ± 1, but has a smaller magnitude, thus reducing the distinguishability. We use a **threshold detector**, which decides that if the correlation is positive, the received pulse is $p(t)$, and if the correlation is negative, the received pulse is $-p(t)$.

Suppose, for example, that $p(t)$ has been transmitted. In the ideal case correlation of this pulse at the receiver would be 1, the maximum possible. Now because of the noise and pulse distortion, the correlation is less than 1. In some extreme situation, channel noise, pulse distortion and overlapping (spreading) from other pulses can make this pulse so dissimilar to $p(t)$ that the correlation can be negative. In this case, the threshold detector decides that $-p(t)$ has been received, thus causing a detection error. In the same way, if $-p(t)$ is transmitted, the detector could decide that $p(t)$ is transmitted. Our task is to make sure that the transmitted pulses have sufficient energy for the damage caused by noise and other imperfections to remain within a limit so that the error probability is below some acceptable bounds. In the ideal case, the margin provided by the correlation c_n for distinguishing the two pulses is 2 (from 1 to -1 and vice versa). The noise and channel distortion reduce this margin. That is why it is important to start with as large a margin as possible. This explains why the antipodal scheme has the best performance in terms of guarding against channel noise and pulse distortion. However, as mentioned earlier, because of some other reasons, schemes such as an orthogonal scheme, where $c_n = 0$, are also used, even when they provide a smaller margin (from 0 to 1 and vice versa) in distinguishing the pulses. Quantitative discussion of correlation in digital signal detection is discussed in chapters 13 and 14.

In later chapters we shall discuss pulse dispersion and pulse distortion during transmission, as well as the calculation of error probability in the presence of noise.

2.6.2 Correlation Functions

Consider the application of correlation to signal detection in a radar, where a signal pulse is transmitted in order to detect a suspected target. If a target is present, the pulse will be reflected by it. If no target is present, there will be no reflected pulse, just a noise. By detecting the presence or absence of the reflected pulse we confirm the presence or absence of a target. By measuring the time delay between the transmitted and the received (reflected) pulses we determine the distance of the target. Let the transmitted and the reflected pulses be denoted by $g(t)$ and $z(t)$, respectively, as shown in Fig. 2.19. If we were to use Eq. (2.45) directly to measure the correlation coefficient c_n, we would obtain

$$c_n = \frac{1}{\sqrt{E_g E_z}} \int_{-\infty}^{\infty} g(t)z(t)\,dt = 0 \tag{2.48}$$

Thus, the correlation is zero because the pulses are disjoint (nonoverlapping in time). The integral (2.48) will yield zero value even when the pulses are identical but with relative time shift. To avoid this difficulty, we compare the transmitted pulse $g(t)$ with the received pulse $z(t)$ shifted by τ. If for some value of τ, there is a strong correlation, we not only detect the presence of the pulse but we also detect the relative time shift of $z(t)$ with respect to $g(t)$. For this reason, instead of using the integral on the right hand, we use the modified integral $\psi_{gz}(\tau)$, the **cross-correlation** function of two real signals $g(t)$ and $z(t)$, defined by*

$$\psi_{gz}(\tau) \equiv \int_{-\infty}^{\infty} g(t)z(t+\tau)\,dt \tag{2.49}$$

Here $z(t+\tau)$ is the pulse $z(t)$ left-shifted (advanced) by τ seconds. Therefore, $\psi_{gz}(\tau)$ is an indication of similarity (correlation) of $g(t)$ with $z(t)$ advanced (left-shifted) by τ seconds.

Autocorrelation Function

The correlation of a signal with itself is called **autocorrelation**. The autocorrelation function $\psi_g(\tau)$ of a real signal $g(t)$ is defined as

$$\psi_g(\tau) \equiv \int_{-\infty}^{\infty} g(t)g(t+\tau)\,dt \tag{2.50}$$

In Chapter 3, we shall show that the autocorrelation function provides a valuable spectral information about the signal.

2.7 SIGNAL REPRESENTATION BY ORTHOGONAL SIGNAL SET

In this section we show a way of representing a signal as a sum of orthogonal signals. Here again we can benefit from the insight gained from a similar problem with vectors. We know that a vector can be represented as the sum of orthogonal vectors, which form the

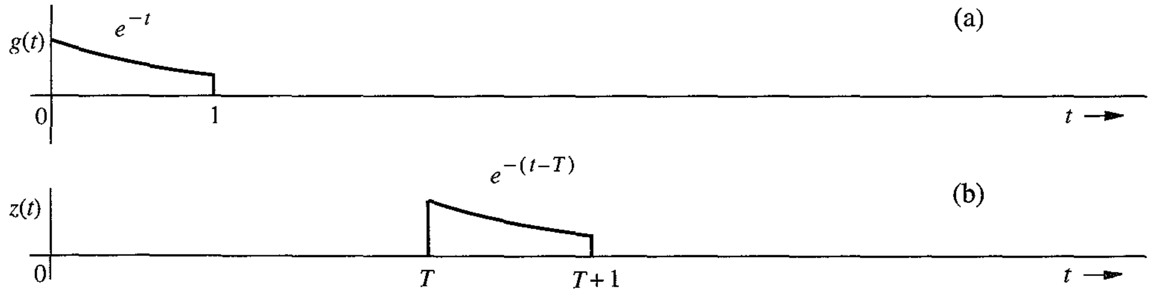


Figure 2.19 Physical explanation of the autocorrelation function.

---

* For complex signals we define

$$\psi_{gz}(\tau) \equiv \int_{-\infty}^{\infty} g^*(t)z(t+\tau)\,dt$$

coordinate system of a vector space. The problem with signals is analogous, and the results for signals are parallel to those for vectors. For this reason let us review the case of vector representation.

2.7.1 Orthogonal Vector Space

Consider a three-dimensional Cartesian vector space described by three mutually orthogonal vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, as shown in Fig. 2.20. First, we shall seek to approximate a three-dimensional vector $\mathbf{g}$ in terms of two mutually orthogonal vectors $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$\mathbf{g} \simeq c_1\mathbf{x}_1 + c_2\mathbf{x}_2$$

The error $\mathbf{e}$ in this approximation is

$$\mathbf{e} = \mathbf{g} - (c_1\mathbf{x}_1 + c_2\mathbf{x}_2)$$

or

$$\mathbf{g} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \mathbf{e}$$

As in the earlier geometrical argument, we see from Fig 2.20 that the length of $\mathbf{e}$ is minimum when $\mathbf{e}$ is perpendicular to the $\mathbf{x}_1$–$\mathbf{x}_2$ plane, and $c_1\mathbf{x}_1$ and $c_2\mathbf{x}_2$ are the projections (components) of $\mathbf{g}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Therefore, the constants c_1 and c_2 are given by Eq. (2.26).

Now let us determine the best approximation to $\mathbf{g}$ in terms of all three mutually orthogonal vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$:

$$\mathbf{g} \simeq c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3 \tag{2.51}$$

Figure 2.20 shows that a unique choice of c_1, c_2, and c_3 exists, for which Eq. (2.51) is no longer an approximation but an equality:

$$\mathbf{g} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3$$

In this case, $c_1\mathbf{x}_1$, $c_2\mathbf{x}_2$, and $c_3\mathbf{x}_3$ are the projections (components) of $\mathbf{g}$ on $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, respectively. Note that the error in the approximation is zero when $\mathbf{g}$ is approximated in terms of three mutually orthogonal vectors: $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. This is because $\mathbf{g}$ is a three-dimensional

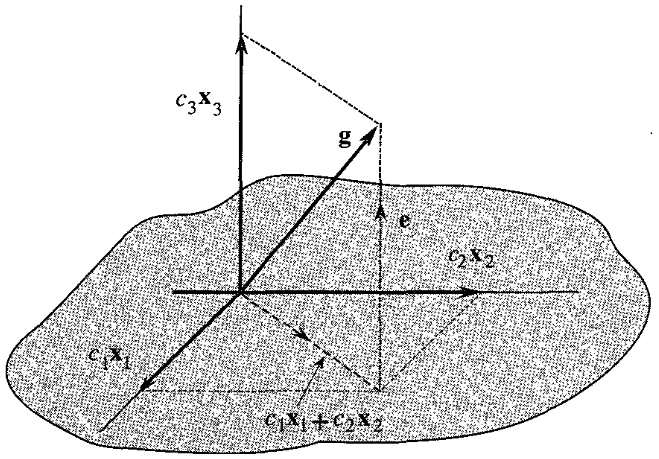


Figure 2.20 Representation of a vector in three-dimensional space.

vector, and the vectors $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ represent a *complete set* of orthogonal vectors in three-dimensional space. Completeness here means that it is impossible to find in this space another vector $\mathbf{x}_4$, that is orthogonal to all three vectors $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$. Any vector in this space can then be represented (with zero error) in terms of these three vectors. Such vectors are known as **basis** vectors. If a set of vectors $\{\mathbf{x}_i\}$ is not complete, the error in the approximation will generally not be zero. Thus, in the three-dimensional case discussed, it is generally not possible to represent a vector $\mathbf{g}$ in terms of only two basis vectors without an error.

The choice of basis vectors is not unique. In fact, a set of basis vectors corresponds to a particular choice of coordinate system. Thus, a three-dimensional vector $\mathbf{g}$ may be represented in many different ways, depending on the coordinate system used.

To summarize, if a set of vectors $\{\mathbf{x}_i\}$ is mutually orthogonal, that is, if

$$\mathbf{x}_m \cdot \mathbf{x}_n = \begin{cases} 0 & m \neq n \\ |\mathbf{x}_m|^2 & m = n \end{cases} \tag{2.52}$$

and if this basis set is complete, a vector $\mathbf{g}$ in this space can be expressed as

$$\mathbf{g} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3 \tag{2.53}$$

where the constants c_i are given by

$$c_i = \frac{\mathbf{g} \cdot \mathbf{x}_i}{\mathbf{x}_i \cdot \mathbf{x}_i} \tag{2.54a}$$

$$= \frac{1}{|\mathbf{x}_i|^2}\,\mathbf{g} \cdot \mathbf{x}_i \qquad i = 1,\ 2,\ 3 \tag{2.54b}$$

2.7.2 Orthogonal Signal Space

We continue with our signal approximation problem using clues and insights developed for vector approximation. As before, we define the orthogonality of a signal set $x_1(t),\ x_2(t),\ \cdots,\ x_N(t)$ over the interval $[t_1,\ t_2]$ as

$$\int_{t_1}^{t_2} x_m(t)x_n^*(t)\,dt = \begin{cases} 0 & m \neq n \\ E_n & m = n \end{cases} \tag{2.55}$$

If the energies $E_n = 1$ for all n, then the set is **normalized** and is called an **orthonormal set**. An orthogonal set can always be normalized by dividing $x_n(t)$ by $\sqrt{E_n}$ for all n. Now, consider the problem of approximating a signal $g(t)$ over the interval $[t_1,\ t_2]$ by a set of N mutually orthogonal signals $x_1(t), x_2(t), \ldots, x_N(t)$:

$$g(t) \simeq c_1x_1(t) + c_2x_2(t) + \cdots + c_Nx_N(t) \tag{2.56a}$$

$$= \sum_{n=1}^{N} c_nx_n(t) \qquad t_1 \leq t \leq t_2 \tag{2.56b}$$

It can be shown that E_e, the energy of the error signal $e(t)$, in this approximation is minimized if we choose[2]

$$c_n = \frac{\int_{t_1}^{t_2} g(t)x_n^*(t)\,dt}{\int_{t_1}^{t_2} {x_n}^2(t)\,dt} \tag{2.57a}$$

$$= \frac{1}{E_n}\int_{t_1}^{t_2} g(t)x_n^*(t)\,dt \qquad n = 1, 2, \ldots, N \tag{2.57b}$$

Moreover, if the orthogonal set is **complete**, the error energy → 0, and the representation in Eqs.(2.56) is no longer an approximation, but an equality,

$$\begin{aligned} g(t) &= c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t) + \cdots \\ &= \sum_{n=1}^{\infty} c_nx_n(t) \qquad t_1 \le t \le t_2 \end{aligned} \tag{2.58}$$

where the coefficients c_n are given by Eq. (2.57). Because the error signal energy approaches zero, it follows that the energy of $g(t)$ is now equal to the sum of the energies of its orthogonal components $c_1x_1(t),\ c_2x_2(t),\ c_3x_3(t),\ \cdots$.

The series on the right-hand side of Eq. (2.58) is called the **generalized Fourier series** of $g(t)$ with respect to the set $\{x_n(t)\}$. When the set $\{x_n(t)\}$ is such that the error energy $E_e \to 0$ as $N \to \infty$ for every member of some particular class, we say that the set $\{x_n(t)\}$ is complete on $[t_1,\ t_2]$ for that class of $g(t)$, and the set $\{x_n(t)\}$ is called a set of **basis functions** or **basis signals**. Unless otherwise mentioned, in the future we shall consider only the class of energy signals.

Thus, when the set $\{x_n(t)\}$ is complete, we have the equality (2.58). One subtle point that must be understood clearly is the meaning of equality in Eq. (2.58). *The equality here is not an equality in the ordinary sense, but in the sense that the error energy, that is, the energy of the difference between the two sides of Eq. (2.58), approaches zero.* If the equality exists in the ordinary sense, the error energy is always zero, but the converse is not necessarily true. The error energy can approach zero even though $e(t)$, the difference between the two sides, is nonzero at some isolated instants. This is because even if $e(t)$ is nonzero at such instants, the area under $e^2(t)$ is still zero. Thus, the Fourier series on the right-hand side of Eq. (2.58) may differ from $g(t)$ at a finite number of points. In fact, when $g(t)$ has a jump discontinuity at $t = t_0$, the corresponding Fourier series at t_0 converges to the mean of $g(t_0{}^+)$ and $g(t_0{}^-)$.

Parseval's Theorem

Recall that the energy of the sum of orthogonal signals is equal to the sum of their energies. Therefore, the energy of the right-hand side of Eq. (2.58) is the sum of the energies of the individual orthogonal components. The energy of a component $c_nx_n(t)$ is $c_n^2E_n$. Equating the energies of the two sides of Eq. (2.58) yields

$$\begin{aligned} E_g &= c_1^2E_1 + c_2^2E_2 + c_3^2E_3 + \cdots \\ &= \sum_n c_n^2E_n \end{aligned} \tag{2.59}$$

This important result is called **Parseval's theorem**. Recall that the signal energy (the area under the squared value of a signal) is analogous to the square of the length of a vector in the vector-signal analogy. In vector space we know that the square of the length of a vector is equal to the sum of the squares of the lengths of its orthogonal components. Parseval's theorem [Eq. (2.59)] is the statement of this fact as applied to signals.

Some Examples of Generalized Fourier Series

The signal representation by generalized Fourier series shows that signals are vectors in every sense. Just as a vector can be represented as a sum of its components in a variety of ways, depending on the choice of a coordinate system, a signal can be represented as a sum of its components in a variety of ways. Just as we have vector coordinate systems formed by mutually orthogonal vectors, such as rectangular, cylindrical, spherical, and so on, we also have signal coordinate systems (basis signals) formed by a variety of sets of mutually orthogonal signals. There exists a large number of orthogonal signal sets which can be used as basis signals for generalized Fourier series. Some well-known signal sets are trigonometric (sinusoid) functions, exponential functions, Walsh functions, Bessel functions, Legendre polynomials, Laguerre functions, Jacobi polynomials, Hermite polynomials, and Chebyshev polynomials. The functions that concern us most in this book are the trigonometric and the exponential sets discussed in the rest of this chapter.

2.8 TRIGONOMETRIC FOURIER SERIES

Consider a signal set:

$$\{1,\ \cos \omega_0 t,\ \cos 2\omega_0 t,\ \ldots,\ \cos n\omega_0 t,\ \ldots, \sin \omega_0 t,\ \sin 2\omega_0 t,\ \ldots,\ \sin n\omega_0 t,\ \ldots\} \tag{2.60}$$

A sinusoid of frequency $n\omega_0$ is called the **nth harmonic** of the sinusoid of frequency ω_0 when n is an integer. The sinusoid of frequency ω_0 serves as an anchor in this set, called the **fundamental**, of which all the remaining terms are harmonics. Note that the constant term 1 is the 0th harmonic in this set because $\cos (0 \times \omega_0 t) = 1$. We can show that this set is orthogonal over any interval of duration $T_0 = 2\pi/\omega_0$, which is the period of the fundamental. This follows from the equations (proved in Appendix A.1)

$$\int_{T_0} \cos n\omega_0 t \ \cos m\omega_0 t \, dt = \begin{cases} 0 & n \neq m \\ \frac{T_0}{2} & m = n \neq 0 \end{cases} \tag{2.61a}$$

$$\int_{T_0} \sin n\omega_0 t \ \sin m\omega_0 t \, dt = \begin{cases} 0 & n \neq m \\ \frac{T_0}{2} & n = m \neq 0 \end{cases} \tag{2.61b}$$

and

$$\int_{T_0} \sin n\omega_0 t \ \cos m\omega_0 t \, dt = 0 \qquad \text{for all } n \text{ and } m \tag{2.61c}$$

The notation $\int_{T_0}$ means integral over an interval from $t = t_1$ to $t = t_1 + T_0$ for any value of t_1. These equations show that the set (2.60) is orthogonal over any contiguous interval of duration T_0. This is the **trigonometric set**, which can be shown to be a complete set.[3, 4] Therefore, we can express a signal $g(t)$ by a trigonometric Fourier series over any interval of duration T_0 seconds as

$$g(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + \cdots + b_1 \sin \omega_0 t + b_2 \sin 2\omega_0 t + \cdots \qquad t_1 \le t \le t_1 + T_0 \tag{2.62a}$$

or

$$g(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + b_n \sin n\omega_0 t \qquad t_1 \le t \le t_1 + T_0 \tag{2.62b}$$

where

$$\omega_0 = \frac{2\pi}{T_0} \tag{2.63}$$

Using Eq. (2.57), we can determine the Fourier coefficients a_0, a_n, and b_n. Thus,

$$a_n = \frac{\displaystyle\int_{t_1}^{t_1+T_0} g(t) \cos n\omega_0 t \, dt}{\displaystyle\int_{t_1}^{t_1+T_0} \cos^2 n\omega_0 t \, dt} \tag{2.64}$$

The integral in the denominator of Eq. (2.64) as seen from Eq. (2.61a) (with $m = n$) is $T_0/2$ when $n \neq 0$. Moreover, for $n = 0$, the denominator is T_0. Hence,

$$a_0 = \frac{1}{T_0} \int_{t_1}^{t_1+T_0} g(t) \, dt \tag{2.65a}$$

and

$$a_n = \frac{2}{T_0} \int_{t_1}^{t_1+T_0} g(t) \cos n\omega_0 t \, dt \qquad n = 1, 2, 3, \ldots \tag{2.65b}$$

Using a similar argument, we obtain

$$b_n = \frac{2}{T_0} \int_{t_1}^{t_1+T_0} g(t) \sin n\omega_0 t \, dt \qquad n = 1, 2, 3, \ldots \tag{2.65c}$$

Compact Trigonometric Fourier Series

The trigonometric Fourier series in Eq. (2.62) contains sine and cosine terms of the same frequency. We can combine the two terms in a single term of the same frequency using the trigonometric identity

$$a_n \cos n\omega_0 t + b_n \sin n\omega_0 t = C_n \cos (n\omega_0 t + \theta_n) \tag{2.66}$$

where

$$C_n = \sqrt{{a_n}^2 + {b_n}^2} \tag{2.67a}$$

$$\theta_n = \tan^{-1} \left(\frac{-b_n}{a_n} \right) \tag{2.67b}$$

For consistency we denote the dc term a_0 by C_0, that is,

$$C_0 = a_0 \tag{2.67c}$$

Using the identity (2.66), the trigonometric Fourier series in Eq. (2.62) can be expressed in the **compact form** of the trigonometric Fourier series as

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \qquad t_1 \le t \le t_1 + T_0 \tag{2.68}$$

where the coefficients C_n and θ_n are computed from a_n and b_n using Eqs. (2.67).

Equation (2.65a) shows that a_0 (or C_0) is the average value of $g(t)$ (averaged over one period). This value can often be determined by inspection of $g(t)$.

EXAMPLE 2.7 Find the compact trigonometric Fourier series for the exponential $e^{-t/2}$ shown in Fig. 2.21a over the interval $0 \le t \le \pi$.

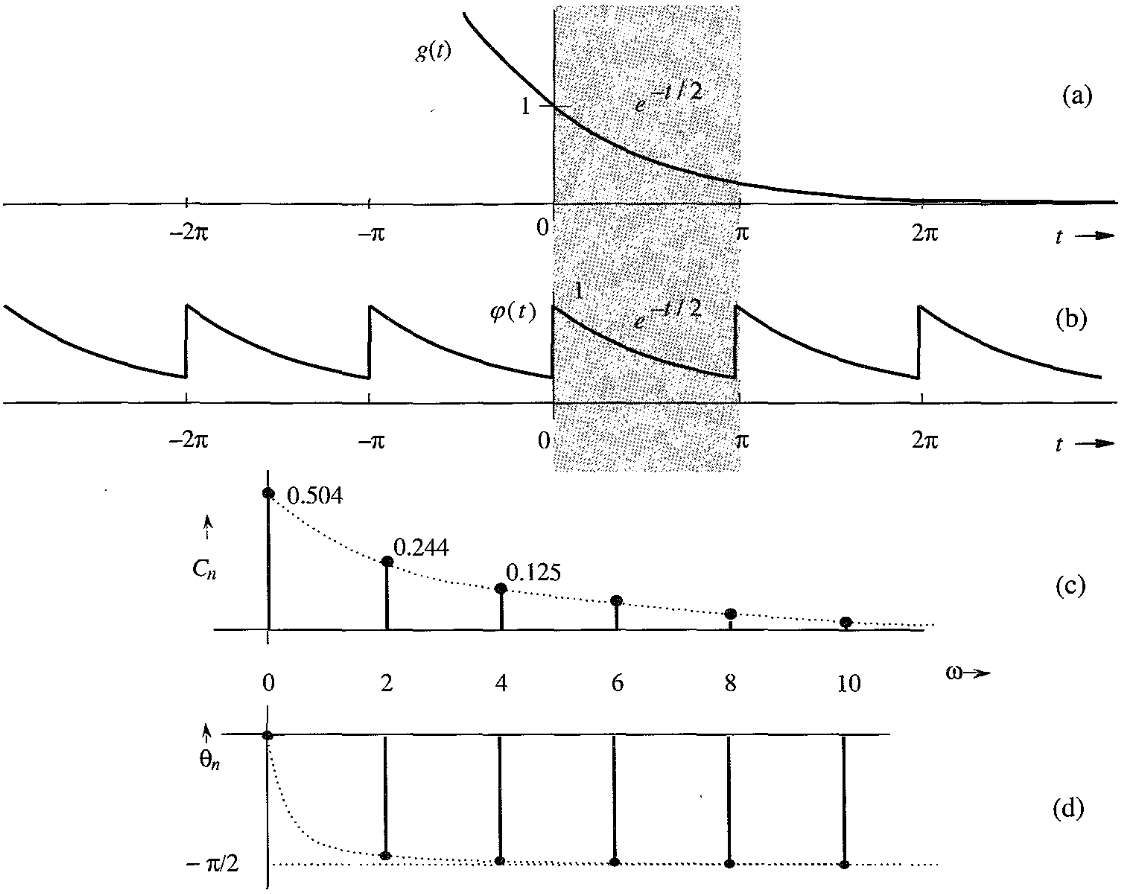


Figure 2.21 Periodic signal and its Fourier spectra.

Because we are required to represent $g(t)$ by the trigonometric Fourier series over the interval $0 \le t \le \pi$, $T_0 = \pi$, and the fundamental frequency is

$$\omega_0 = \frac{2\pi}{T_0} = 2 \text{ rad/s}$$

Therefore,

$$g(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos 2nt + b_n \sin 2nt \qquad 0 \le t \le \pi$$

where [from Eq. (2.65a)]

$$a_0 = \frac{1}{\pi}\int_0^{\pi} e^{-t/2}\,dt = 0.50$$

$$a_n = \frac{2}{\pi}\int_0^{\pi} e^{-t/2}\cos 2nt\,dt = 0.504\left(\frac{2}{1+16n^2}\right)$$

and

$$b_n = \frac{2}{\pi}\int_0^{\pi} e^{-t/2}\sin 2nt\,dt = 0.504\left(\frac{8n}{1+16n^2}\right)$$

Therefore,

$$g(t) = 0.504\left[1+\sum_{n=1}^{\infty}\frac{2}{1+16n^2}(\cos 2nt + 4n\sin 2nt)\right] \qquad 0 \le t \le \pi$$

To find the compact Fourier series, we find its coefficients using Eqs. (2.67) as

$$C_0 = a_0 = 0.504$$

$$C_n = \sqrt{a_n^2+b_n^2} = 0.504\sqrt{\frac{4}{(1+16n^2)^2}+\frac{64n^2}{(1+16n^2)^2}} = 0.504\left(\frac{2}{\sqrt{1+16n^2}}\right)$$

$$\theta_n = \tan^{-1}\left(\frac{-b_n}{a_n}\right) = \tan^{-1}(-4n) = -\tan^{-1}4n \tag{2.69}$$

The amplitudes and phases of the dc and the first seven harmonics are computed from Eq. (2.69) and displayed in Table 2.1. Using these numerical values, we can express $g(t)$ in the compact trigonometric Fourier series as

$$g(t) = 0.504 + 0.504\sum_{n=1}^{\infty}\frac{2}{\sqrt{1+16n^2}}\cos(2nt - \tan^{-1}4n) \qquad 0 \le t \le \pi \tag{2.70a}$$

$$= 0.504 + 0.244\cos(2t - 75.96°) + 0.125\cos(4t - 82.87°) + 0.084\cos(6t - 85.24°) + 0.063\cos(8t - 86.42°) + \cdots \quad 0 \le t \le \pi \tag{2.70b}$$

Table 2.1

n	0	1	2	3	4	5	6	7
C_n	0.504	0.244	0.125	0.084	0.063	0.0504	0.042	0.036
θ_n	0	−75.96	−82.87	−85.24	−86.42	−87.14	−87.61	−87.95

Periodicity of the Trigonometric Fourier Series

We have shown how an arbitrary signal $g(t)$ may be expressed as a trigonometric Fourier series over any interval of T_0 seconds. The Fourier series is equal to $g(t)$ over this interval alone. Outside this interval the series is not necessarily equal to $g(t)$. It would be interesting to find out what happens to the Fourier series outside this interval. We now show that the trigonometric Fourier series is a periodic function of period T_0 (the period of the fundamental).

Let us denote the trigonometric Fourier series on the right-hand side of Eq. (2.68) by $\varphi(t)$. Therefore,

$$\varphi(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos (n\omega_0 t + \theta_n) \qquad \text{for all } t$$

and

$$\begin{aligned}\varphi(t + T_0) &= C_0 + \sum_{n=1}^{\infty} C_n \cos [n\omega_0(t + T_0) + \theta_n] \\ &= C_0 + \sum_{n=1}^{\infty} C_n \cos (n\omega_0 t + 2n\pi + \theta_n) \\ &= C_0 + \sum_{n=1}^{\infty} C_n \cos (n\omega_0 t + \theta_n) \\ &= \varphi(t) \qquad \text{for all } t \end{aligned} \tag{2.71}$$

This shows that the trigonometric Fourier series is a periodic function of period T_0 (the period of its fundamental). For instance, $\varphi(t)$, the Fourier series on the right-hand side of Eq. (2.70), is a periodic function in which the segment of $g(t)$ in Fig. 2.21a over the interval $0 \leq t \leq \pi$ repeats periodically every π seconds, as shown in Fig. 2.21b.* Thus, when we represent a signal $g(t)$ by the trigonometric Fourier series over a certain interval of duration T_0, the function $g(t)$ and its Fourier series $\varphi(t)$ need only be equal over that interval of T_0 seconds. Outside this interval, the Fourier series repeats periodically with period T_0. Now if the function $g(t)$ were itself to be periodic with period T_0, then a Fourier series representing $g(t)$ over an interval T_0 will also represent $g(t)$ for all t (not just over the interval T_0). Moreover, such a periodic signal $g(t)$ can be generated by a periodic repetition of any of its segments of duration T_0 (see Sec. 2.2.3, Fig. 2.7). Therefore, the trigonometric Fourier series representing a segment of $g(t)$ of duration T_0 starting at any instant represents $g(t)$ for all t. This means in computing the coefficients a_0, a_n, and b_n, we may use any value for t_1 in Eqs. (2.65). In other words, we may perform this integration over any interval of T_0. Thus, the Fourier coefficients of a series representing a periodic signal $g(t)$ (for all t) can be expressed as

$$a_0 = \frac{1}{T_0} \int_{T_0} g(t)\, dt \tag{2.72a}$$

$$a_n = \frac{2}{T_0} \int_{T_0} g(t) \cos n\omega_0 t\, dt \qquad n = 1, 2, 3, \dots \tag{2.72b}$$

and

$$b_n = \frac{2}{T_0} \int_{T_0} g(t) \sin n\omega_0 t\, dt \qquad n = 1, 2, 3, \dots \tag{2.72c}$$

where $\int_{T_0}$ means that the integration is performed over any interval of T_0 seconds.

---

* In reality, the series convergence at the points of discontinuity shows about 9% overshoot (Gibbs phenomenon).[2]

Fourier Spectrum

The compact trigonometric Fourier series in Eq. (2.68) indicates that a periodic signal $g(t)$ can be expressed as a sum of sinusoids of frequencies 0 (dc), $\omega_0, 2\omega_0, \cdots, n\omega_0, \cdots$, whose amplitudes are $C_0, C_1, C_2, \ldots, C_n, \cdots$ and whose phases are $0, \theta_1, \theta_2, \cdots, \theta_n, \cdots$. We can readily plot amplitude C_n vs. ω (**amplitude spectrum**) and θ_n vs. ω (**phase spectrum**). These two plots together are the **frequency spectra** of $g(t)$.

Figure 2.21c and d show the amplitude and phase spectra for the periodic signal $\varphi(t)$ in Fig. 2.21b. These spectra tell us at a glance the frequency composition of $\varphi(t)$, that is, the amplitudes and phases of various sinusoidal components of $\varphi(t)$. Knowing the frequency spectra, we can reconstruct or synthesize $\varphi(t)$, as shown on the right-hand side of Eq. (2.70). Therefore, the frequency spectra in Fig. 2.21c and d provide an alternative description—the **frequency-domain description** of $\varphi(t)$. The **time-domain description** of $\varphi(t)$ is shown in Fig. 2.21b. *A signal, therefore, has a dual identity: the time-domain identity $\varphi(t)$ and the frequency-domain identity (Fourier spectra). The two identities complement each other. Taken together, they provide a better understanding of a signal.*

Series Convergence at Jump Discontinuities

When there is a jump discontinuity in a periodic signal $g(t)$, its Fourier series at the point of discontinuity converges to an average of the left-hand and right-hand limits of $g(t)$ at the instant of discontinuity*. In Fig. 2.21b, for instance, the periodic signal $\varphi(t)$ is discontinuous at $t = 0$ with $\varphi(0^+) = 1$ and $\varphi(0^-) = e^{-\pi/2} = 0.208$. The corresponding Fourier series converges to a value of $(1 + 0.208)/2 = 0.604$ at $t = 0$. This is easily verified from Eq. (2.70b) by setting $t = 0$.

Existence of the Fourier Series: Dirichlet Conditions

There are two basic conditions for the existence of the Fourier series.

1. For the series to exist, the coefficients a_0, a_n, and b_n in Eqs. (2.65) must be finite. From Eqs. (2.65) it follows that the existence of these coefficients is guaranteed if $g(t)$ is absolutely integrable over one period; that is,

$$\int_{T_0} |g(t)|\, dt < \infty \tag{2.73}$$

This is known as the **weak Dirichlet condition.** *If a function $g(t)$ satisfies the weak Dirichlet condition, the existence of a Fourier series is guaranteed, but the series may not converge at every point.* For example, if a function $g(t)$ is infinite at some point, then obviously the series representing the function will be nonconvergent at that point. Similarly, if a function has an infinite number of maxima and minima in one period, then the function contains an appreciable amount of components of frequencies approaching infinity. Thus, the higher coefficients in the series do not decay rapidly, so that the series will not converge rapidly or uniformly. Thus, for a convergent Fourier series, in addition to condition (2.73), we require that:

2. The function $g(t)$ have only a finite number of maxima and minima in one period, and it may have only a finite number of finite discontinuities in one period.

These two conditions are known as the **strong Dirichlet conditions**. We note here that any periodic waveform that can be generated in a laboratory satisfies strong Dirichlet conditions, and hence possesses a convergent Fourier series. Thus, a physical possibility of a periodic waveform is a valid and sufficient condition for the existence of a convergent series.

---

* This behavior of the Fourier series is dictated by its error energy minimization property, discussed in Sec. 2.7.

EXAMPLE 2.8 Find the compact trigonometric Fourier series for the periodic square wave $w(t)$ shown in Fig. 2.22a, and sketch its amplitude and phase spectra.

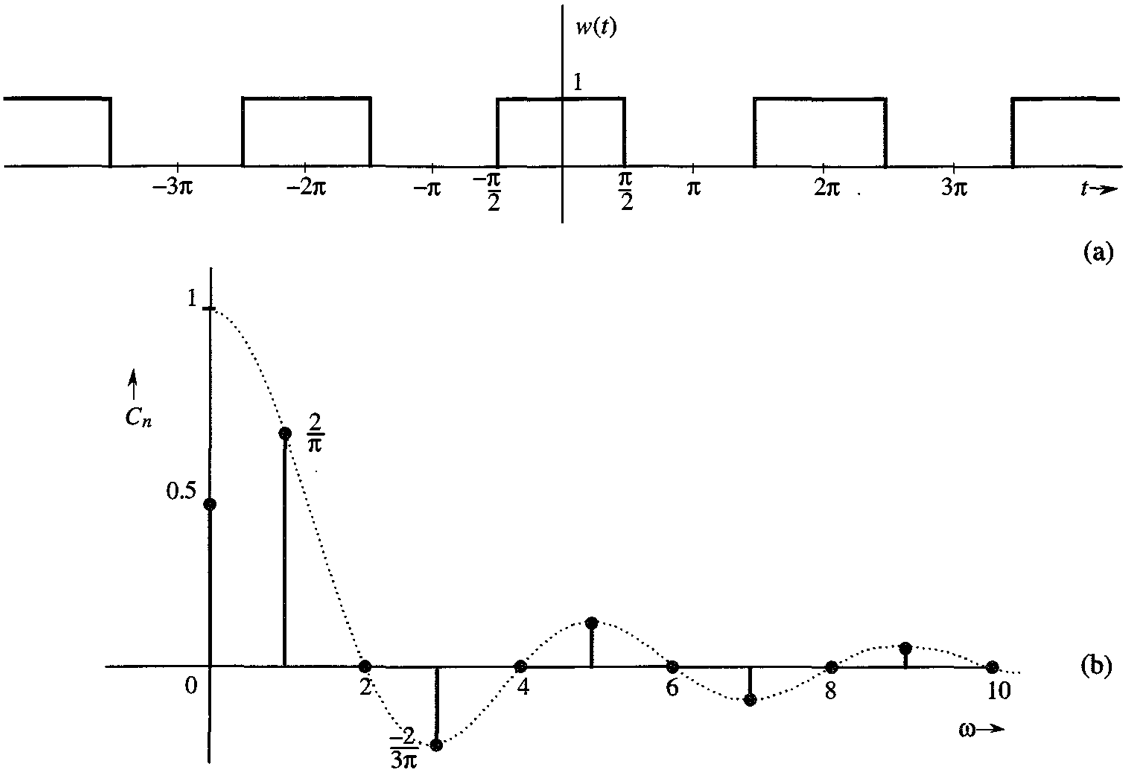


Figure 2.22 Square pulse periodic signal and its Fourier spectra.

The Fourier series is

$$w(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + b_n \sin n\omega_0 t$$

where

$$a_0 = \frac{1}{T_0} \int_{T_0} w(t)\, dt$$

In the preceding equation, we may integrate $w(t)$ over any interval of duration T_0. Figure 2.22a shows that the best choice for a region of integration is from $-T_0/2$ to $T_0/2$. Because $w(t) = 1$ only over $(-T_0/4,\ T_0/4)$ and $w(t) = 0$ over the remaining segment,

$$a_0 = \frac{1}{T_0} \int_{-T_0/4}^{T_0/4} dt = \frac{1}{2} \tag{2.74a}$$

We could have found a_0, the average value of $w(t)$, to be $1/2$ merely by inspection of $w(t)$ in Fig. 2.22a. Also,

$$a_n = \frac{2}{T_0}\int_{-T_0/4}^{T_0/4} \cos n\omega_0 t \, dt = \frac{2}{n\pi} \sin\left(\frac{n\pi}{2}\right)$$

$$= \begin{cases} 0 & n \text{ even} \\ \dfrac{2}{\pi n} & n = 1, 5, 9, 13, \cdots \\ -\dfrac{2}{\pi n} & n = 3, 7, 11, 15, \cdots \end{cases} \tag{2.74b}$$

$$b_n = \frac{2}{T_0}\int_{-T_0/4}^{T_0/4} \sin nt \, dt = 0 \tag{2.74c}$$

In these derivations we used the fact that $\omega_0 T_0 = 2\pi$. Therefore,

$$w(t) = \frac{1}{2} + \frac{2}{\pi}\left(\cos \omega_0 t - \frac{1}{3}\cos 3\omega_0 t + \frac{1}{5}\cos 5\omega_0 t - \frac{1}{7}\cos 7\omega_0 t + \cdots\right) \tag{2.75}$$

Observe that $b_n = 0$ and all the sine terms are zero. Only the cosine terms appear in the trigonometric series. The series is therefore already in compact form, except that the amplitudes of alternating harmonics are negative. Now by definition, amplitudes C_n are positive [see Eq. (2.67a)]. The negative sign can be accommodated by a phase of π radians. This can be seen from the trigonometric identity*

$$-\cos x = \cos(x - \pi)$$

Using this fact, we can express the series in Eq. (2.75) as

$$w(t) = \frac{1}{2} + \frac{2}{\pi}\left[\cos \omega_0 t + \frac{1}{3}\cos(3\omega_0 t - \pi) + \frac{1}{5}\cos 5\omega_0 t \right.$$
$$\left. + \frac{1}{7}\cos(7\omega_0 t - \pi) + \frac{1}{9}\cos 9\omega_0 t + \cdots\right]$$

This is the desired form of the compact trigonometric Fourier series. The amplitudes are

$$C_0 = \frac{1}{2}$$

$$C_n = \begin{cases} 0 & n \text{ even} \\ \dfrac{2}{\pi n} & n \text{ odd} \end{cases}$$

$$\theta_n = \begin{cases} 0 & \text{for all } n \neq 3, 7, 11, 15, \cdots \\ -\pi & n = 3, 7, 11, 15, \cdots \end{cases}$$

We could plot amplitude and phase spectra using these values. We can, however, simplify our task in this special case if we allow amplitude C_n to take on negative values. If this is allowed, we do not need a phase of $-\pi$ to account for the sign. This means the phases of all components are zero, and we can discard the phase spectrum and manage with only the amplitude spectrum, as shown in Fig. 2.22b. Observe that there is no loss of information in doing so and that the amplitude spectrum in Fig. 2.22b has the complete information about the Fourier series in Eq. (2.75). Therefore, *whenever all sine terms vanish ($b_n = 0$), it is convenient to allow C_n to take on negative values.* This permits the spectral information to be conveyed by a single spectrum—the amplitude spectrum. Because C_n can be positive as well as negative, the spectrum is called the **amplitude spectrum** rather than the magnitude spectrum.

---

* Because $\cos(x \pm \pi) = -\cos x$, we could have chosen the phase π or $-\pi$. In fact, $\cos(x \pm N\pi) = -\cos x$ for any odd integral value of N. Therefore, the phase can be chosen as $\pm N\pi$, where N is any convenient odd integer.

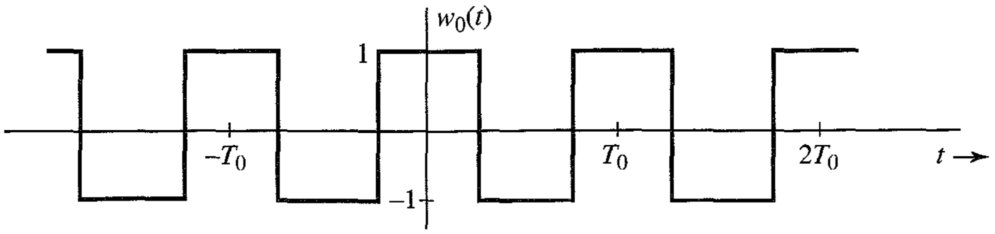


Figure 2.23 Bipolar square pulse periodic signal.

Another useful function that is related to the periodic square wave is the bipolar square wave $w_0(t)$ shown in Fig. 2.23a. We encounter this signal in switching applications. Note that $w_0(t)$ is basically $w(t)$ minus its dc component. It is easy to see that

$$w_0(t) = 2[w(t) - 0.5]$$

Hence, from Eq. (2.75) it follows that

$$w_0(t) = \frac{4}{\pi}\left(\cos \omega_0 t - \frac{1}{3}\cos 3\omega_0 t + \frac{1}{5}\cos 5\omega_0 t - \frac{1}{7}\cos 7\omega_0 t + \cdots\right) \quad (2.76)$$

Comparison of this equation with Eq. (2.75) shows that the Fourier components of $w_0(t)$ are identical to those of $w(t)$ [Eq. (2.75)] in every respect except for doubling the amplitudes and loss of dc.

EXAMPLE 2.9 Find the trigonometric Fourier series and sketch the corresponding spectra for the periodic impulse train $\delta_{T_0}(t)$ shown in Fig. 2.24a.

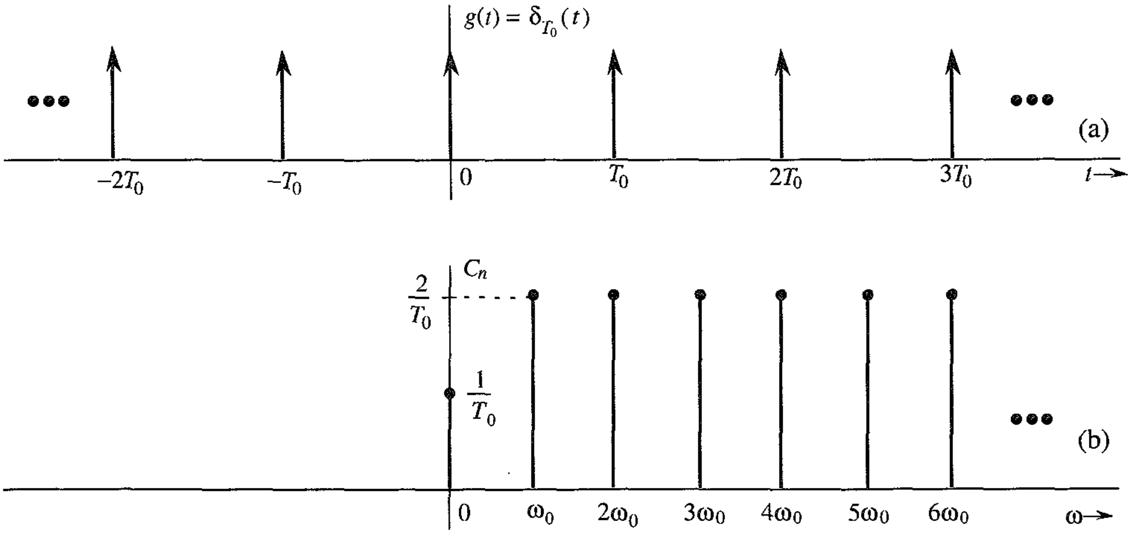


Figure 2.24 Impulse train and its Fourier spectrum.

The trigonometric Fourier series for $\delta_{T_0}(t)$ is given by

$$\delta_{T_0}(t) = C_0 + \sum C_n \cos(n\omega_0 t + \theta_n) \qquad \omega_0 = \frac{2\pi}{T_0}$$

We first compute a_0, a_n, and b_n:

$$a_0 = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} \delta(t)\, dt = \frac{1}{T_0}$$

$$a_n = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2} \delta(t) \cos n\omega_0 t\, dt = \frac{2}{T_0}$$

This result follows from the sampling property (2.19) of the impulse function. Similarly, using the sampling property of the impulse, we obtain

$$b_n = \frac{2}{T_0}\int_{-T_0/2}^{T_0/2} \delta(t) \sin n\omega_0 t\, dt = 0$$

Therefore, $C_0 = 1/T_0$, $C_n = 2/T_0$, and $\theta_n = 0$. Thus,

$$\delta_{T_0}(t) = \frac{1}{T_0}\left(1 + 2\sum_{n=1}^{\infty} \cos n\omega_0 t\right) \tag{2.77}$$

Figure 2.24b shows the amplitude spectrum. The phase spectrum is zero.

Effect of Symmetry

The Fourier series for the signal $g(t)$ in Fig. 2.21a (Example 2.7) consists of sine and cosine terms, but the series for the signal $w(t)$ in Fig. 2.22a (Example 2.8) consists of cosine terms only. In some cases the Fourier series consists of sine terms only. This is no accident. It can be shown that the Fourier series of any even periodic function $g(t)$ consists of cosine terms only and the series of any odd periodic function $g(t)$ consists of sine terms only (see Prob. 2.8-3).

2.9 EXPONENTIAL FOURIER SERIES

It is shown in Appendix A.2 that the set of exponentials $e^{jn\omega_0 t}$ $(n = 0, \pm1, \pm2, \ldots)$ is orthogonal over any interval of duration $T_0 = 2\pi/\omega_0$; that is,

$$\int_{T_0} e^{jm\omega_0 t}(e^{jn\omega_0 t})^*\, dt = \int_{T_0} e^{j(m-n)\omega_0 t}\, dt = \begin{cases} 0 & m \neq n \\ T_0 & m = n \end{cases} \tag{2.78}$$

Moreover, this set is a complete set.[3, 4] From Eqs. (2.58) and (2.57) it follows that a signal $g(t)$ can be expressed over an interval of duration T_0 seconds as an exponential Fourier series

$$g(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \tag{2.79}$$

where [see Eq. (2.57)]

$$D_n = \frac{1}{T_0}\int_{T_0} g(t) e^{-jn\omega_0 t}\, dt \tag{2.80}$$

The exponential Fourier series is basically another form of the trigonometric Fourier series. Each sinusoid of frequency ω can be expressed as the sum of the two exponentials $e^{j\omega t}$ and $e^{-j\omega t}$. This results in the exponential Fourier series consisting of components of the form $e^{jn\omega_0 t}$ with n varying from $-\infty$ to ∞. The exponential Fourier series in Eq. (2.79) is periodic with period T_0.

In order to see its close connection with the trigonometric series, we shall rederive the exponential Fourier series from the trigonometric Fourier series. A sinusoid in the trigonometric series can be expressed as a sum of two exponentials using Euler's formula:

$$\begin{aligned} C_n \cos(n\omega_0 t + \theta_n) &= \frac{C_n}{2}\left[e^{j(n\omega_0 t+\theta_n)} + e^{-j(n\omega_0 t+\theta_n)}\right] \\ &= \left(\frac{C_n}{2}e^{j\theta_n}\right)e^{jn\omega_0 t} + \left(\frac{C_n}{2}e^{-j\theta_n}\right)e^{-jn\omega_0 t} \\ &= D_n e^{jn\omega_0 t} + D_{-n}e^{-jn\omega_0 t} \end{aligned} \tag{2.81}$$

where

$$\begin{aligned} D_n &= \frac{1}{2}C_n e^{j\theta_n} \\ D_{-n} &= \frac{1}{2}C_n e^{-j\theta_n} \end{aligned} \tag{2.82}$$

The compact trigonometric Fourier series of a periodic signal $g(t)$ is given by

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

The use of Eq. (2.81) in the preceding equation (and letting $C_0 = D_0$) yields

$$\begin{aligned} g(t) &= D_0 + \sum_{n=1}^{\infty} D_n e^{jn\omega_0 t} + D_{-n}e^{-jn\omega_0 t} \\ &= D_0 + \sum_{n=-\infty\ (n\neq 0)}^{\infty} D_n e^{jn\omega_0 t} \end{aligned}$$

which is precisely equivalent to Eq. (2.79) derived earlier. Equations (2.82) show the close connection between the coefficients of the trigonometric and the exponential Fourier series.

Observe the compactness of expressions (2.79) and (2.80) and compare them to expressions corresponding to trigonometric Fourier series. These two equations demonstrate very clearly the principal virtue of exponential Fourier series. First, the form of the series is more compact. Second, the mathematical expression for deriving the coefficients of the series is also compact. It is much more convenient to handle the exponential series than the trigonometric one. In the system analysis also, the exponential form proves more convenient than the trigonometric form. For these reasons we shall use exponential (rather than trigonometric) representation of signals in the rest of the book.

EXAMPLE 2.10 Find the exponential Fourier series for the signal in Fig. 2.21b (Example 2.7).

In this case, $T_0 = \pi$, $\omega_0 = 2\pi/T_0 = 2$, and

$$\varphi(t) = \sum_{n=-\infty}^{\infty} D_n e^{j2nt}$$

where

$$\begin{aligned} D_n &= \frac{1}{T_0}\int_{T_0} \varphi(t)e^{-j2nt}\,dt \\ &= \frac{1}{\pi}\int_0^{\pi} e^{-t/2}\,e^{-j2nt}\,dt \\ &= \frac{1}{\pi}\int_0^{\pi} e^{-(\frac{1}{2}+j2n)t}\,dt \\ &= \frac{-1}{\pi\left(\frac{1}{2}+j2n\right)} e^{-(\frac{1}{2}+j2n)t}\Bigg|_0^{\pi} \\ &= \frac{0.504}{1+j4n} \end{aligned} \tag{2.83}$$

and

$$\varphi(t) = 0.504 \sum_{n=-\infty}^{\infty} \frac{1}{1+j4n} e^{j2nt} \tag{2.84a}$$

$$\begin{aligned} &= 0.504\left[1 + \frac{1}{1+j4}e^{j2t} + \frac{1}{1+j8}e^{j4t} + \frac{1}{1+j12}e^{j6t} + \cdots \right. \\ &\quad \left. + \frac{1}{1-j4}e^{-j2t} + \frac{1}{1-j8}e^{-j4t} + \frac{1}{1-j12}e^{-j6t} + \cdots \right] \end{aligned} \tag{2.84b}$$

Observe that the coefficients D_n are complex. Moreover, D_n and D_{-n} are conjugates, as expected [see Eqs. (2.82)].

2.9.1 Exponential Fourier Spectra

In exponential spectra, we plot coefficients D_n as a function of ω. But since D_n is complex in general, we need two plots: the real and the imaginary parts of D_n or the amplitude (magnitude) and the angle of D_n. We prefer the latter because of its close connection to the amplitudes and phases of corresponding components of the trigonometric Fourier series. We therefore plot $|D_n|$ vs. ω and $\angle D_n$ vs. ω. This requires that the coefficients D_n be expressed in polar form as $|D_n|e^{j\angle D_n}$.

A comparison of Eqs. (2.65a) and (2.80) (for $n = 0$) shows that $D_0 = a_0 = C_0$. Equations (2.82) show that for a real periodic signal the twin coefficients D_n and D_{-n} are conjugates, and

$$|D_n| = |D_{-n}| = \frac{1}{2}C_n \tag{2.85a}$$

$$\angle D_n = \theta_n \qquad \text{and} \qquad \angle D_{-n} = -\theta_n \tag{2.85b}$$

Thus,

$$D_n = |D_n|e^{j\theta_n} \quad \text{and} \quad D_{-n} = |D_n|e^{-j\theta_n} \tag{2.86}$$

Note that $|D_n|$ are the amplitudes (magnitudes) and $\angle D_n$ are the angles of various exponential components. From Eqs. (2.85) it follows that the amplitude spectrum ($|D_n|$ vs. ω) is an even function of ω and the angle spectrum ($\angle D_n$ vs. ω) is an odd function of ω when $g(t)$ is a real signal.

For the series in Example 2.10 [Eq. (2.84b)], for instance,

$$\begin{aligned}
D_0 &= 0.504 \\
D_1 &= \frac{0.504}{1+j4} = 0.122e^{-j75.96^\circ} \Longrightarrow |D_1| = 0.122 \qquad \angle D_1 = -75.96^\circ \\
D_{-1} &= \frac{0.504}{1-j4} = 0.122e^{j75.96^\circ} \Longrightarrow |D_{-1}| = 0.122 \qquad \angle D_{-1} = 75.96^\circ
\end{aligned}$$

and

$$\begin{aligned}
D_2 &= \frac{0.504}{1+j8} = 0.0625e^{-j82.87^\circ} \Longrightarrow |D_2| = 0.0625 \qquad \angle D_2 = -82.87^\circ \\
D_{-2} &= \frac{0.504}{1-j8} = 0.0625e^{j82.87^\circ} \Longrightarrow |D_{-2}| = 0.0625 \qquad \angle D_{-2} = 82.87^\circ
\end{aligned}$$

and so on. Note that D_n and D_{-n} are conjugates, as expected [see Eqs. (2.85)].

Figure 2.25 shows the frequency spectra (amplitude and angle) of the exponential Fourier series for the periodic signal $\varphi(t)$ in Fig. 2.21b.

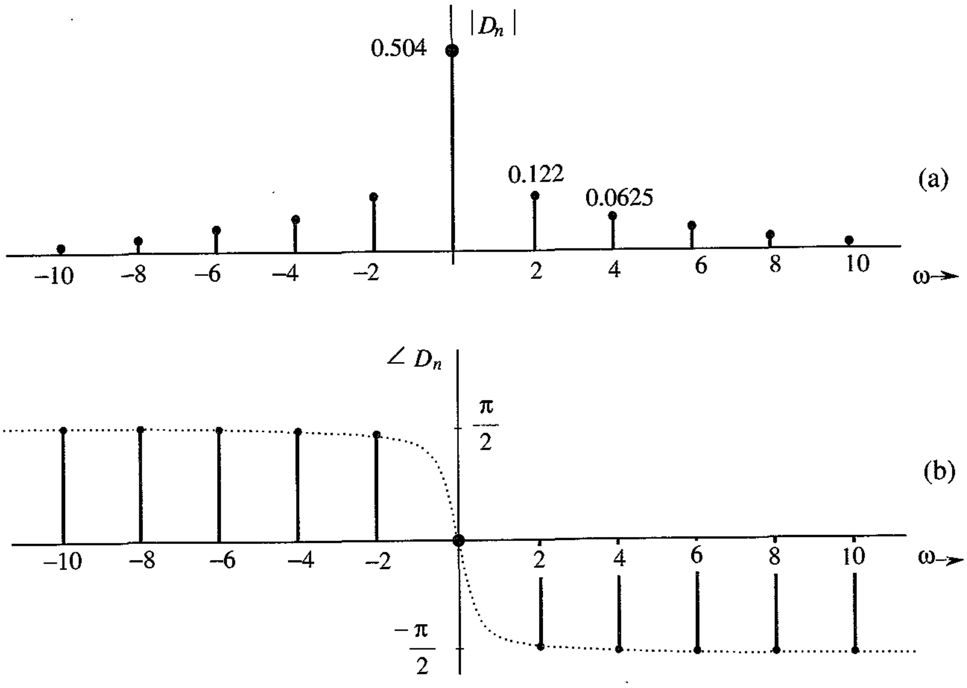


Figure 2.25 Exponential Fourier spectra for the signal in Fig. 2.21a.

We notice some interesting features of these spectra. First, the spectra exist for positive as well as negative values of ω (the frequency). Second, the amplitude spectrum is an even function of ω and the angle spectrum is an odd function of ω. Finally, we see a close connection between these spectra and the spectra of the corresponding trigonometric Fourier series for $\varphi(t)$ (Fig. 2.21c and d).

What Is a Negative Frequency?

The existence of the spectrum at negative frequencies is somewhat disturbing because by definition, the frequency (number of repetitions per second) is a positive quantity. How do we interpret a negative frequency? Using a trigonometric identity, the sinusoid of a negative frequency $-\omega_0$ can be expressed as

$$\cos(-\omega_0 t + \theta) = \cos(\omega_0 t - \theta)$$

This clearly shows that the frequency of a sinusoid $\cos(\omega_0 t + \theta)$ is $|\omega_0|$, which is a positive quantity. The same conclusion is reached by observing that

$$e^{\pm j\omega_0 t} = \cos \omega_0 t \pm j \sin \omega_0 t$$

Thus, the frequency of exponentials $e^{\pm j\omega_0 t}$ is indeed $|\omega_0|$. How do we then interpret the spectral plots for negative values of ω? A healthier way of looking at the situation is to say that *exponential spectra are a graphical representation of coefficients D_n as a function of ω. Existence of the spectrum at $\omega = -n\omega_0$ is merely an indication of the fact that an exponential component $e^{-jn\omega_0 t}$ exists in the series.* We know that a sinusoid of frequency $n\omega_0$ can be expressed in terms of a pair of exponentials $e^{jn\omega_0 t}$ and $e^{-jn\omega_0 t}$ [see Eq. (2.81)].

Equations (2.85) show the close connection between trigonometric spectra (C_n and θ_n) and exponential spectra ($|D_n|$ and $\angle D_n$). The dc components D_0 and C_0 are identical in both spectra. Moreover, the exponential amplitude spectrum $|D_n|$ is half the trigonometric amplitude spectrum C_n for $n \geq 1$. The exponential angle spectrum $\angle D_n$ is identical to the trigonometric phase spectrum θ_n for $n \geq 0$. We can therefore produce the exponential spectra merely by the inspection of trigonometric spectra, and vice versa.

EXAMPLE 2.11 Find the exponential Fourier series for the periodic square wave $w(t)$ shown in Fig. 2.22a.

We have

$$w(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t}$$

where

$$\begin{aligned} D_n &= \frac{1}{T_0}\int_{T_0} w(t)e^{-jn\omega_0 t}\,dt \\ &= \frac{1}{T_0}\int_{-T_0/4}^{T_0/4} e^{-jn\omega_0 t}\,dt \\ &= \frac{1}{-jn\omega_0 T_0}\left(e^{-jn\omega_0 T_0/4} - e^{jn\omega_0 T_0/4}\right) \\ &= \frac{2}{n\omega_0 T_0}\sin\left(\frac{n\omega_0 T_0}{4}\right) = \frac{1}{n\pi}\sin\left(\frac{n\pi}{2}\right) \end{aligned}$$

In this case D_n is real. Consequently, we can do without the phase or angle plot if we plot D_n vs ω instead of the amplitude spectrum ($|D_n|$ vs. ω), as shown in Fig. 2.26. Compare this spectrum with the trigonometric spectrum in Fig. 2.22b. Observe that $D_0 = C_0$ and $|D_n| = |D_{-n}| = C_n/2$, as expected.

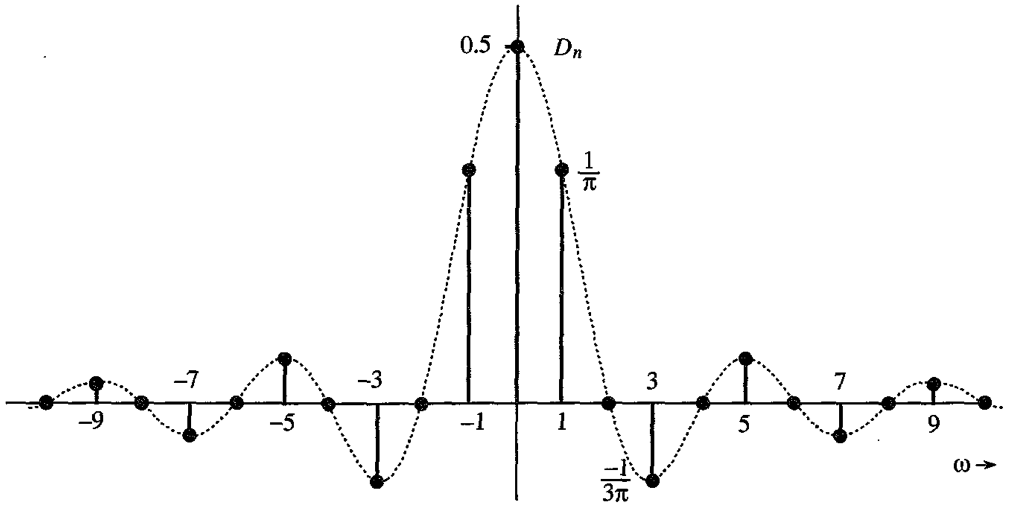


Figure 2.26 Exponential Fourier spectrum of the square pulse periodic signal.

EXAMPLE 2.12 Find the exponential Fourier series and sketch the corresponding spectra for the impulse train $\delta_{T_0}(t)$ shown in Fig. 2.27.

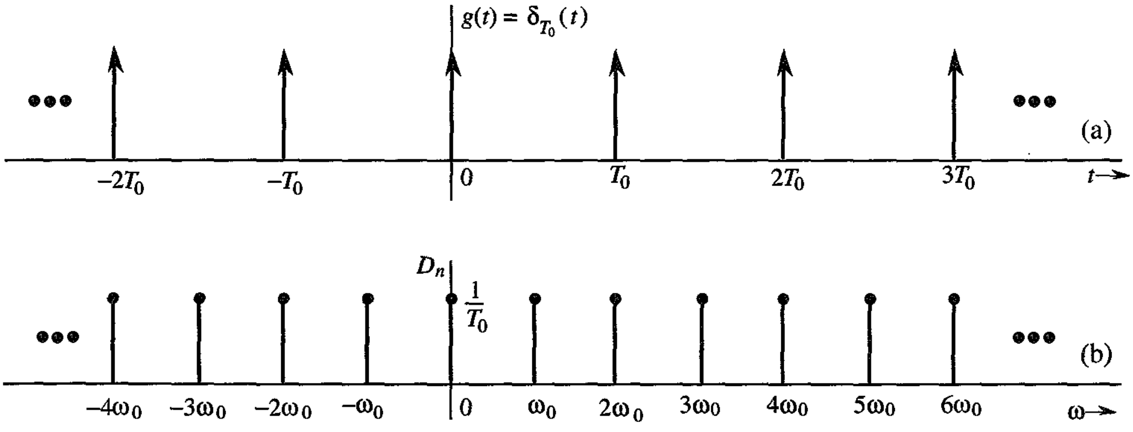


Figure 2.27 Impulse train and its exponential Fourier spectra.

The exponential Fourier series is given by

$$\delta_{T_0}(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \qquad \omega_0 = \frac{2\pi}{T_0} \tag{2.87}$$

where

$$D_n = \frac{1}{T_0} \int_{T_0} \delta_{T_0}(t) e^{-jn\omega_0 t}\, dt$$

Choosing the interval of integration $(-T_0/2,\ T_0/2)$ and recognizing that over this interval $\delta_{T_0}(t) = \delta(t)$,

$$D_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} \delta(t) e^{-jn\omega_0 t}\, dt$$

In this integral, the impulse is located at $t = 0$. From the sampling property of the impulse function, the integral on the right-hand side is the value of $e^{-jn\omega_0 t}$ at $t = 0$ (where the impulse is located). Therefore,

$$D_n = \frac{1}{T_0} \tag{2.88}$$

and

$$\delta_{T_0}(t) = \frac{1}{T_0} \sum_{n=-\infty}^{\infty} e^{jn\omega_0 t} \qquad \omega_0 = \frac{2\pi}{T_0} \tag{2.89}$$

Equation (2.89) shows that the exponential spectrum is uniform ($D_n = 1/T_0$) for all the frequencies, as shown in Fig. 2.27. The spectrum, being real, requires only the amplitude plot. All phases are zero. Compare this spectrum to the trigonometric spectrum shown in Fig. 2.24b. The dc components are identical and the exponential spectrum amplitudes are half those in the trigonometric spectrum for all $\omega > 0$.

Parseval's Theorem

A periodic signal $g(t)$ is a power signal, and every term in its Fourier series is also a power signal. The power P_g of $g(t)$ is equal to the power of its Fourier series. Because the Fourier series consists of terms that are mutually orthogonal over one period, the power of the Fourier series is equal to the sum of the powers of its Fourier components. This follows from Parseval's theorem. We have already demonstrated this result in Example 2.2 for the trigonometric Fourier series. It is also valid for the exponential Fourier series. Thus, for the trigonometric Fourier series

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

the power of $g(t)$ is given by

$$P_g = C_0{}^2 + \frac{1}{2} \sum_{n=1}^{\infty} C_n{}^2 \tag{2.90}$$

For the exponential Fourier series

$$g(t) = D_0 + \sum_{\substack{n=-\infty \\ (n \neq 0)}}^{\infty} D_n e^{jn\omega_0 t}$$

the power is given by (see Prob. 2.1-7)

$$P_g = \sum_{n=-\infty}^{\infty} |D_n|^2 \tag{2.91a}$$

For a real $g(t)$, $|D_{-n}| = |D_n|$. Therefore,

$$P_g = {D_0}^2 + 2\sum_{n=1}^{\infty} |D_n|^2 \tag{2.91b}$$

Comments: Parseval's theorem occurs in many different forms, such as in Eqs. (2.59), (2.90), and (2.91). Yet another form is found in Eq. (3.64). Although these forms appear different, they all state the same principle, that is, the square of the length of a vector equals the sum of the squares of its orthogonal components. The form (2.59) applies to energy signals, the form (2.90) applies to periodic signals represented by the trigonometric Fourier series, and the form (2.91) applies to periodic signals represented by the exponential Fourier series.

2.10 NUMERICAL COMPUTATION OF D_n

We can compute D_n numerically using the **discrete Fourier transform (DFT)**, which uses the samples of a periodic signal $g(t)$ over one period. The sampling interval is T_s seconds. Hence, there are $N_0 = T_0/T_s$ number of samples in one period T_0. To find the relationship between D_n and the samples of $g(t)$, consider Eq. (2.80),

$$\begin{aligned} D_n &= \frac{1}{T_0}\int_{T_0} g(t)e^{-jn\omega_0 t}\,dt \\ &= \lim_{T_s\to 0} \frac{1}{T_0}\sum_{k=0}^{N_0-1} g(kT_s)e^{-jn\omega_0 kT_s}\,T_s \\ &= \lim_{T_s\to 0} \frac{1}{N_0}\sum_{k=0}^{N_0-1} g(kT_s)e^{-jn\Omega_0 k} \end{aligned} \tag{2.92}$$

where $g(kT_s)$ is the kth sample of $g(t)$ and

$$\Omega_0 = \omega_0 T_s, \qquad N_0 = \frac{T_0}{T_s} \tag{2.93}$$

In practice, it is impossible to make $T_s \to 0$ in computing the right-hand side of Eq. (2.92). We can make it small, but not zero because it will increase the data without limit. Thus, we shall ignore the limit on T_s in Eq. (2.92) with the implicit understanding that T_s is reasonably small. This results in some computational error, which is inevitable in any numerical evaluation of an integral. The error resulting from nonzero T_s is called the **aliasing error**, which will be discussed in more details in Chapter 6. Thus, we can express Eq. (2.92) as

$$D_n = \frac{1}{N_0}\sum_{k=0}^{N_0-1} g(kT_s)e^{-jn\Omega_0 k} \tag{2.94}$$

This equation shows that $D_{n+N_0} = D_n$. Hence, Eq. (2.94) yields the Fourier spectrum D_n repeating periodically with period N_0. This will result in overlapping of various components. To reduce the effect of such overlapping, we need to increase N_0 as much as practicable. We shall see later [Sec. (6.1)] that the overlapping appears as if the spectrum above the $(N_0/2)$th harmonics had folded back at this frequency $(N_0\omega_0/2)$. Hence, to minimize the effect of this spectral folding, we should make sure that D_n for $n \geq N_0/2$ is negligible. The DFT (or FFT) gives the coefficients D_n for $n \geq 0$ up to $n = N_0/2$. Beyond $n = N_0/2$, the coefficients represent the values for negative n because of the periodicity property $D_{n+N_0} = D_n$. For instance, when $N_0 = 32$, $D_{17} = D_{-15}$, $D_{18} = D_{-14}$, $\cdots$, $D_{31} = D_{-1}$. The cycle repeats again from $n = 32$ on.

We can use the efficient **fast Fourier transform (FFT)** to compute the right-hand side of Eq. (2.94). We shall use MATLAB to implement the FFT algorithm. For this purpose, we need samples of $g(t)$ over one period starting at $t = 0$. In this algorithm, it is also preferable (although not necessary) that N_0 be a power of 2, that is, $N_0 = 2^m$, where m is an integer.

Computer Example C2.1

Compute and plot the trigonometric and exponential Fourier spectra for the periodic signal in Fig. 2.21b (Example 2.7).

The samples of $g(t)$ start at $t = 0$, and the last (N_0th) sample is at $t = T_0 - T_s$. (The last sample is not at $t = T_0$ because the sample at $t = 0$ is identical to the sample at $t = T_0$, and the next cycle begins at $t = T_0$.) At the points of discontinuity, the sample value is taken as the average of the values of the function on two sides of the discontinuity. Thus, in the present case, the first sample (at $t = 0$) is not 1, but $(e^{-\pi/2} + 1)/2 = 0.604$. To determine N_0, we require D_n for $n \geq N_0/2$ to be relatively small. Because $g(t)$ has a jump discontinuity, D_n decays rather slowly as $1/n$. Hence, a choice of $N_0 = 200$ is acceptable because the $(N_0/2)$th (100th) harmonic is about 0.01 (about 1%) of the fundamental. However, we also require N_0 to be a power of 2. Hence, we shall take $N_0 = 256 = 2^8$.

We write and save a MATLAB file (or program) c21.m to compute and plot the Fourier coefficients.

```
% (c21.m)
%M is the number of coefficients to be computed
T0=pi;N0=256;Ts=T0/N0;M=10;
t=0:Ts:Ts*(N0-1); t=t';
g=exp(-t/2);g(1)=0.604;
% fft(g) is the FFT [the sum on the right-hand side of Eq. (2.94)]
Dn=fft(g)/N0
[Dnangle,Dnmag]=cart2pol(real(Dn),imag(Dn));
k=0:length(Dn)-1;k=k';
subplot(211),stem(k,Dnmag)
subplot(212),stem(k,Dnangle)
```

To compute trigonometric Fourier series coefficients, we recall program c21.m along with commands to convert D_n into C_n and θ_n.

```
c21;clg
C0=Dnmag(1); Cn=2*Dnmag(2:M);
```

```
Amplitudes=[C0;Cn]
Angles=Dnangles(1:M);
Angles=Angles*(180/pi);
disp('Amplitudes Angles')
[Amplitudes Angles]
% To Plot the Fourier coefficients
k=0:length(Amplitudes)-1; k=k';
subplot(211),stem(k,Amplitudes)
subplot(212),stem(k,Angles)
ans =
   Amplitudes   Angles
   0.5043       0
   0.2446       -75.9622
   0.1251       -82.8719
   0.0837       -85.2317
   0.0629       -86.4175
   0.0503       -87.1299
   0.0419       -87.6048
   0.0359       -87.9437
   0.0314       -88.1977
   0.0279       -88.3949
```

REFERENCES

1. A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, New York, 1962.
2. B. P. Lathi, *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, Carmichael, CA, 1998.
3. P. L. Walker, *The Theory of Fourier Series and Integrals*, Wiley-Interscience, New York, 1986.
4. R. V. Churchill, and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed., McGraw-Hill, New York, 1978.

PROBLEMS

2.1-1 Find the energies of the signals shown in Fig. P2.1-1. Comment on the effect on energy of sign change, time shifting or doubling of the signal. What is the effect on the energy if the signal is multiplied by k?

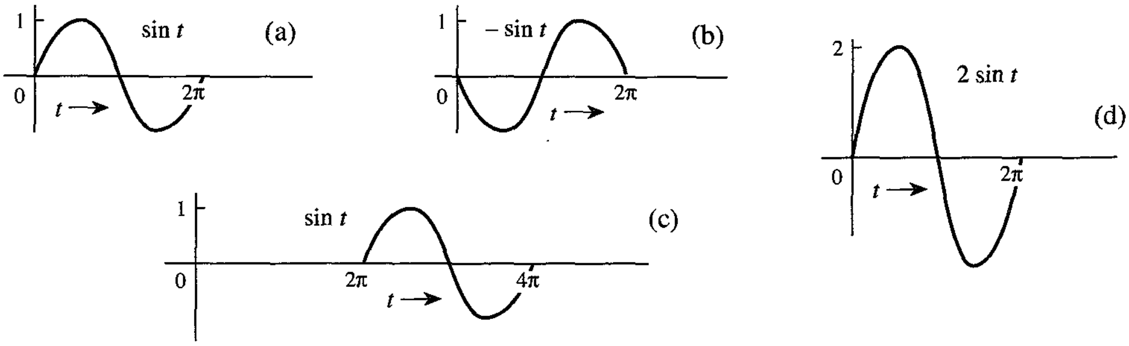


Figure P2.1-1

2.1-2 **(a)** Find E_x and E_y, the energies of the signals $x(t)$ and $y(t)$ shown in Fig. P2.1-2a. Sketch the signals $x(t) + y(t)$ and $x(t) - y(t)$ and show that the energies of either of these two signals are equal to $E_x + E_y$. Repeat the procedure for the signal pair of Fig. P2.1-2b.

(b) Repeat the procedure for the signal pair of Fig. P2.1-2c. Are the energies of the signals $x(t) + y(t)$ and $x(t) - y(t)$ identical in this case?

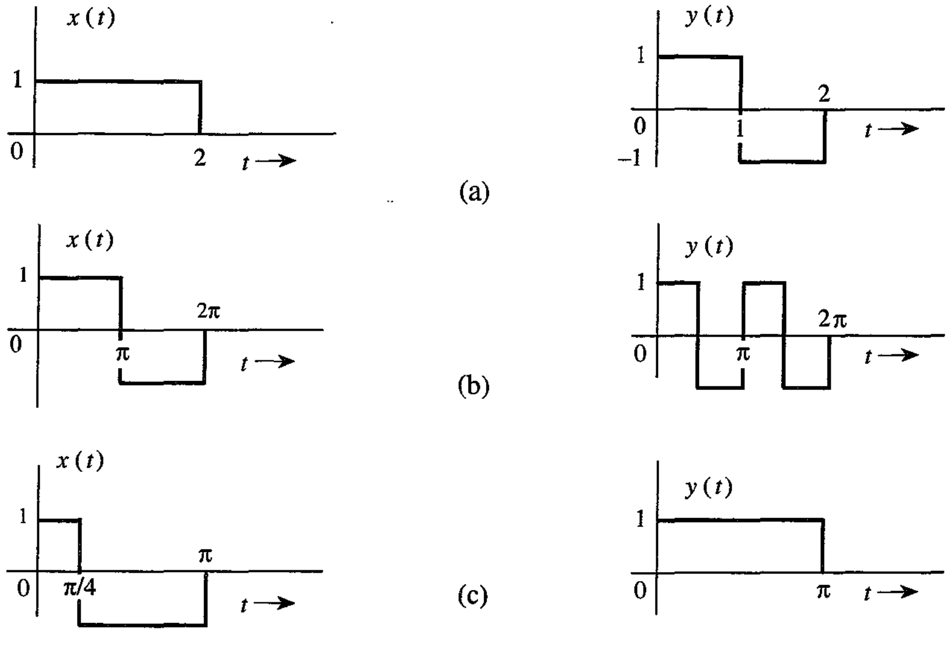


Figure P2.1-2

2.1-3 Redo Example 2.2a to find the power of a sinusoid $C \cos(\omega_0 t + \theta)$ by averaging the signal energy over one period $2\pi/\omega_0$ (rather than averaging over the infinitely large interval).

2.1-4 Show that if $\omega_1 = \omega_2$, the power of $g(t) = C_1 \cos(\omega_1 t + \theta_1) + C_2 \cos(\omega_2 t + \theta_2)$ is $[C_1{}^2 + C_2{}^2 + 2C_1C_2 \cos(\theta_1 - \theta_2)]/2$, which is not equal to $(C_1{}^2 + C_2{}^2)/2$.

2.1-5 Find the power of the periodic signal $g(t)$ shown in Fig. P2.1-5. Find also the powers and the rms values of: **(a)** $-g(t)$; **(b)** $2g(t)$; **(c)** $cg(t)$. Comment.

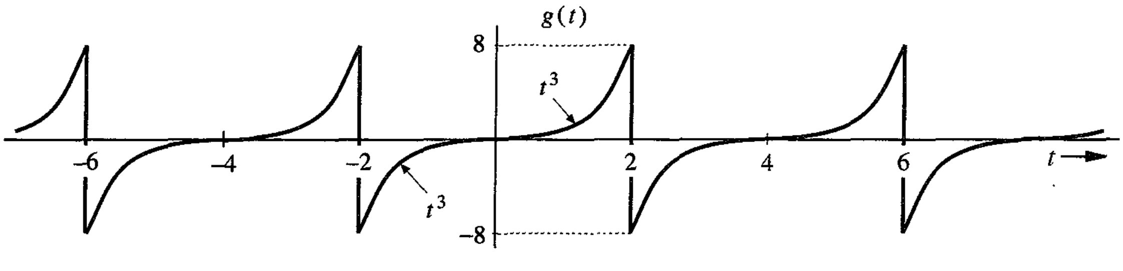


Figure P2.1-5

2.1-6 Find the power and the rms value for the signals in: **(a)** Fig. 2.21b; **(b)** Fig. 2.22a; **(c)** Fig. 2.23; **(d)** Fig. P2.8-4a; **(e)** Fig. P2.8-4c.

2.1-7 Show that the power of a signal $g(t)$ given by

$$g(t) = \sum_{k=m}^{n} D_k e^{j\omega_k t} \qquad \omega_i \neq \omega_k \text{ for all } i \neq k$$

is (Parseval's theorem)

$$P_g = \sum_{k=m}^{n} |D_k|^2$$

2.1-8 Determine the power and the rms value for each of the following signals:

(a) $10 \cos\left(100t + \frac{\pi}{3}\right)$ (b) $10 \cos\left(100t + \frac{\pi}{3}\right) + 16 \sin\left(150t + \frac{\pi}{5}\right)$

(c) $(10 + 2 \sin 3t) \cos 10t$ (d) $10 \cos 5t \cos 10t$

(e) $10 \sin 5t \cos 10t$ (f) $e^{j\alpha t} \cos \omega_0 t$

2.2-1 Show that an exponential e^{-at} starting at $-\infty$ is neither an energy nor a power signal for any real value of a. However, if a is imaginary, it is a power signal with power $P_g = 1$ regardless of the value of a.

2.3-1 In Fig. P2.3-1, the signal $g_1(t) = g(-t)$. Express signals $g_2(t)$, $g_3(t)$, $g_4(t)$, and $g_5(t)$ in terms of signals $g(t)$, $g_1(t)$, and their time-shifted, time-scaled, or time-inverted versions. For instance $g_2(t) = g(t - T) + g_1(t - T)$ for some suitable value of T. Similarly, both $g_3(t)$ and $g_4(t)$ can be expressed as $g(t - T) + g(t + T)$ for some suitable value of T. $g_5(t)$ can be expressed as $g(t)$ time-shifted, time-scaled, and then multiplied by a constant. (These operations may be performed in any order).

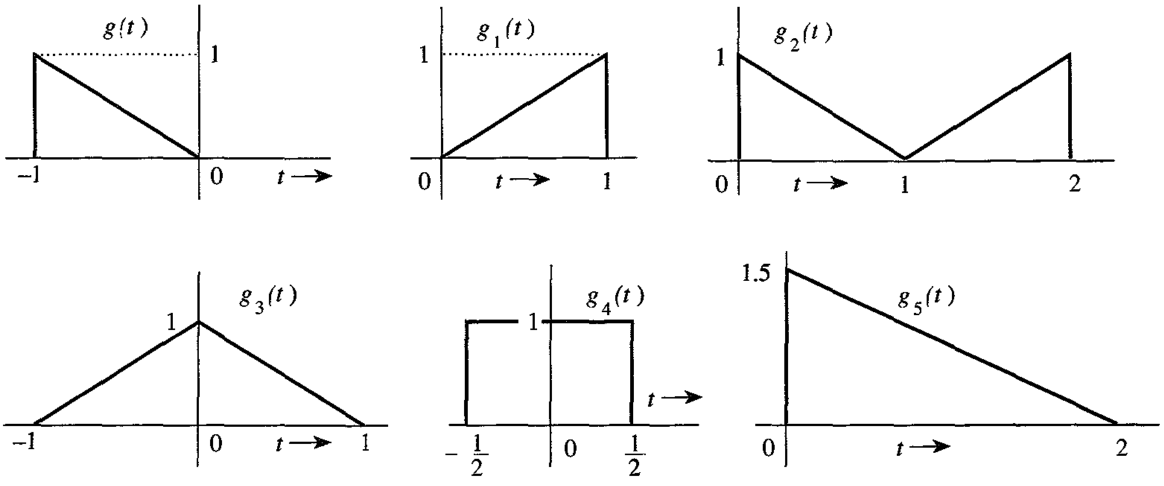


Figure P2.3-1

2.3-2 For the signal $g(t)$ shown in Fig. P2.3-2, sketch the signals: **(a)** $g(-t)$; **(b)** $g(t + 6)$; **(c)** $g(3t)$; **(d)** $g(6 - t)$.

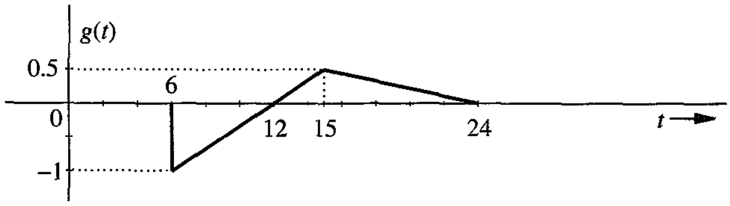


Figure P2.3-2

2.3-3 For the signal $g(t)$ shown in Fig. P2.3-3, sketch: **(a)** $g(t-4)$; **(b)** $g(t/1.5)$; **(c)** $g(2t-4)$ **(d)** $g(2-t)$. *Hint*: Recall that replacing t with $t-T$ delays the signal by T. Thus, $g(2t-4)$ is $g(2t)$ with t replaced by $t-2$. Similarly, $g(2-t)$ is $g(-t)$ with t replaced by $t-2$.

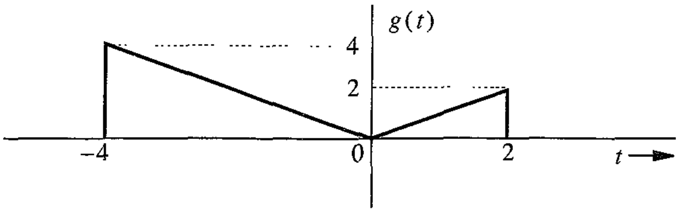


Figure P2.3-3

2.3-4 For an energy signal $g(t)$ with energy E_g, show that the energy of any one of the signals $-g(t)$, $g(-t)$, and $g(t-T)$ is E_g. Show also that the energy of $g(at)$ as well as $g(at-b)$ is E_g/a. This shows that time inversion and time shifting do not affect signal energy. On the other hand, time compression of a signal by a factor a reduces the energy by the factor a. What is the effect on signal energy if the signal is: **(a)** time-expanded by a factor a $(a > 1)$; **(b)** multiplied by a constant a?

2.4-1 Simplify the following expressions:

(a) $\left(\dfrac{\sin t}{t^2+2}\right)\delta(t)$ **(b)** $\left(\dfrac{j\omega+2}{\omega^2+9}\right)\delta(\omega)$

(c) $\left[e^{-t}\cos(3t-60^\circ)\right]\delta(t)$ **(d)** $\left[\dfrac{\sin\frac{\pi}{2}(t-2)}{t^2+4}\right]\delta(t-1)$

(e) $\left(\dfrac{1}{j\omega+2}\right)\delta(\omega+3)$ **(f)** $\left(\dfrac{\sin k\omega}{\omega}\right)\delta(\omega)$

Hint: Use Eq. (2.18). For part **(f)** use L'Hôpital's rule.

2.4-2 Evaluate the following integrals:

(a) $\displaystyle\int_{-\infty}^{\infty} g(\tau)\delta(t-\tau)\,d\tau$ **(b)** $\displaystyle\int_{-\infty}^{\infty} \delta(\tau)g(t-\tau)\,d\tau$

(c) $\displaystyle\int_{-\infty}^{\infty} \delta(t)e^{-j\omega t}\,dt$ **(d)** $\displaystyle\int_{-\infty}^{\infty} \delta(t-2)\sin\pi t\,dt$

(e) $\displaystyle\int_{-\infty}^{\infty} \delta(t+3)e^{-t}\,dt$ **(f)** $\displaystyle\int_{-\infty}^{\infty} (t^3+4)\delta(1-t)\,dt$

(g) $\displaystyle\int_{-\infty}^{\infty} g(2-t)\delta(3-t)\,dt$ **(h)** $\displaystyle\int_{-\infty}^{\infty} e^{(x-1)}\cos\frac{\pi}{2}(x-5)\delta(x-3)\,dx$

Hint: $\delta(x)$ is located at $x=0$. For example, $\delta(1-t)$ is located at $1-t=0$; that is, at $t=1$, and so on.

2.4-3 Prove that

$$\delta(at) = \frac{1}{|a|}\delta(t)$$

Hence, show that

$$\delta(\omega) = \frac{1}{2\pi}\delta(f) \qquad \text{where} \quad \omega = 2\pi f$$

Hint: Show that

$$\int_{-\infty}^{\infty} \phi(t)\delta(at)\, dt = \frac{1}{|a|}\phi(0)$$

2.5-1 Derive Eq. (2.26) in an alternate way by observing that $\mathbf{e} = (\mathbf{g} - c\mathbf{x})$, and

$$|\mathbf{e}|^2 = (\mathbf{g} - c\mathbf{x})\cdot(\mathbf{g} - c\mathbf{x}) = |\mathbf{g}|^2 + c^2|\mathbf{x}|^2 - 2c\mathbf{g}\cdot\mathbf{x}$$

2.5-2 For the signals $g(t)$ and $x(t)$ shown in Fig. P2.5-2, find the component of the form $x(t)$ contained in $g(t)$. In other words, find the optimum value of c in the approximation $g(t) \approx cx(t)$ so that the error signal energy is minimum. What is the error signal energy?

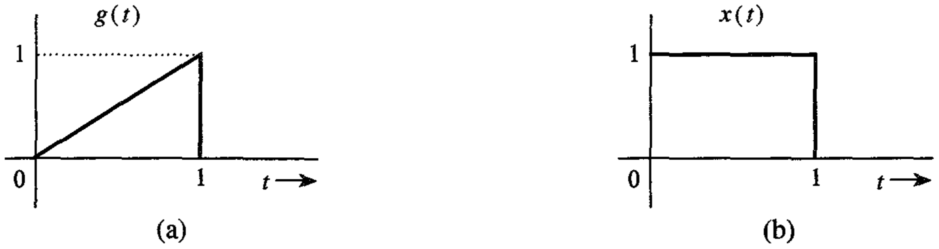


Figure P2.5-2

2.5-3 For the signals $g(t)$ and $x(t)$ shown in Fig. P2.5-2, find the component of the form $g(t)$ contained in $x(t)$. In other words, find the optimum value of c in the approximation $x(t) \approx cg(t)$ so that the error signal energy is minimum. What is the error signal energy?

2.5-4 Repeat Prob. 2.5-2 if $x(t)$ is the sinusoid pulse shown in Fig. P2.5-4.

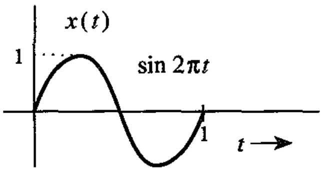


Figure P2.5-4

2.5-5 Energies of the two energy signals $x(t)$ and $y(t)$ are E_x and E_y, respectively.

(a) If $x(t)$ and $y(t)$ are orthogonal, then show that the energy of the signal $x(t) + y(t)$ is identical to the energy of the signal $x(t) - y(t)$, and is given by $E_x + E_y$.

(b) If $x(t)$ and $y(t)$ are orthogonal, find the energies of signals $c_1x(t) + c_2y(t)$ and $c_1x(t) - c_2y(t)$.

(c) We define E_{xy}, the cross energy of the two energy signals $x(t)$ and $y(t)$, as

$$E_{xy} = \int_{-\infty}^{\infty} x(t)y^*(t)\, dt$$

If $z(t) = x(t) \pm y(t)$, then show that

$$E_z = E_x + E_y \pm (E_{xy} + E_{yx})$$

2.5-6 Let $x_1(t)$ and $x_2(t)$ be two unit energy signals orthogonal over an interval from $t = t_1$ to t_2. We can represent $x_1(t)$ and $x_2(t)$ by two unit length, orthogonal vectors ($\mathbf{x}_1$, $\mathbf{x}_2$). Consider a signal $g(t)$ where

$$g(t) = c_1 x_1(t) + c_2 x_2(t) \qquad t_1 \le t \le t_2$$

This signal can be represented as a vector $\mathbf{g}$ by a point $(c_1,\ c_2)$ in the x_1–x_2 plane.

(a) Determine the vector representation of the following six signals in this two-dimensional vector space:

(i) $g_1(t) = 2x_1(t) - x_2(t)$ **(ii)** $g_2(t) = -x_1(t) + 2x_2(t)$
(iii) $g_3(t) = -x_2(t)$ **(iv)** $g_4(t) = x_1(t) + 2x_2(t)$
(v) $g_5(t) = 2x_1(t) + x_2(t)$ **(vi)** $g_6(t) = 3x_1(t)$

(b) Point out pairs of mutually orthogonal vectors among these six vectors. Verify that the pairs of signals corresponding to these orthogonal vectors are also orthogonal.

2.6-1 Find the correlation coefficient c_n of signal $x(t)$ and each of the four pulses $g_1(t)$, $g_2(t)$, $g_3(t)$, and $g_4(t)$ shown in Fig. P2.6-1. Which pair of pulses would you select for a binary communication in order to provide maximum margin against the noise along the transmission path?

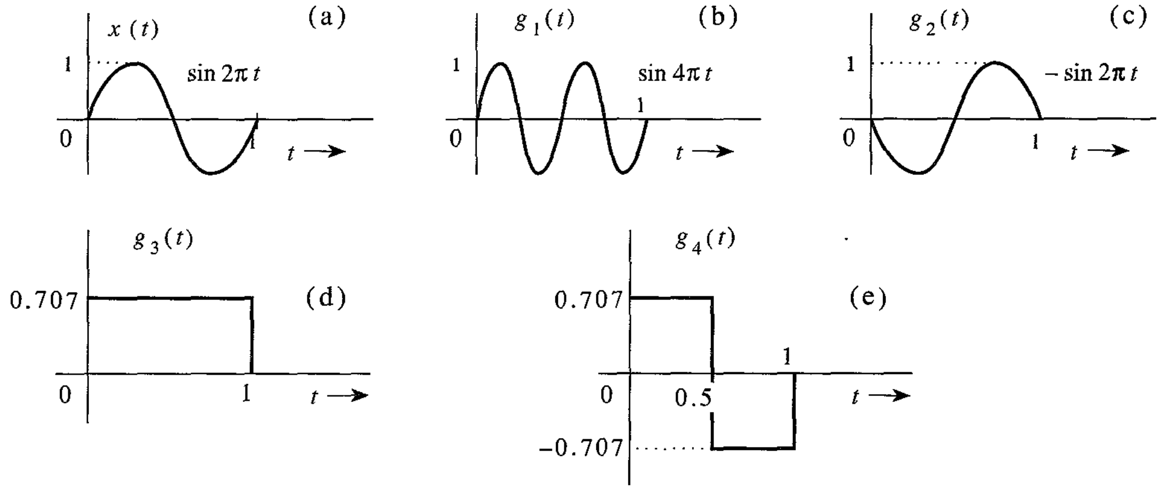

$x(t)$
(a)
1
0
$\sin 2\pi t$
1
$t \rightarrow$
$g_1(t)$
(b)
1
0
$\sin 4\pi t$
1
$t \rightarrow$
$g_2(t)$
(c)
1
0
$-\sin 2\pi t$
1
$t \rightarrow$
$g_3(t)$
0.707
(d)
0
1
$t \rightarrow$
$g_4(t)$
0.707
(e)
1
0
0.5
$t \rightarrow$
-0.707

Figure P2.6-1

2.8-1 **(a)** Sketch the signal $g(t) = t^2$ and find the trigonometric Fourier series to represent $g(t)$ over the interval $(-1,\ 1)$. Sketch the Fourier series $\varphi(t)$ for all values of t.

(b) Verify Parseval's theorem [Eq. (2.90)] for this case, given that

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}$$

2.8-2 (**a**) Sketch the signal $g(t) = t$ and find the trigonometric Fourier series to represent $g(t)$ over the interval $(-\pi,\ \pi)$. Sketch the Fourier series $\varphi(t)$ for all values of t.

(**b**) Verify Parseval's theorem [Eq. (2.90)] for this case, given that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

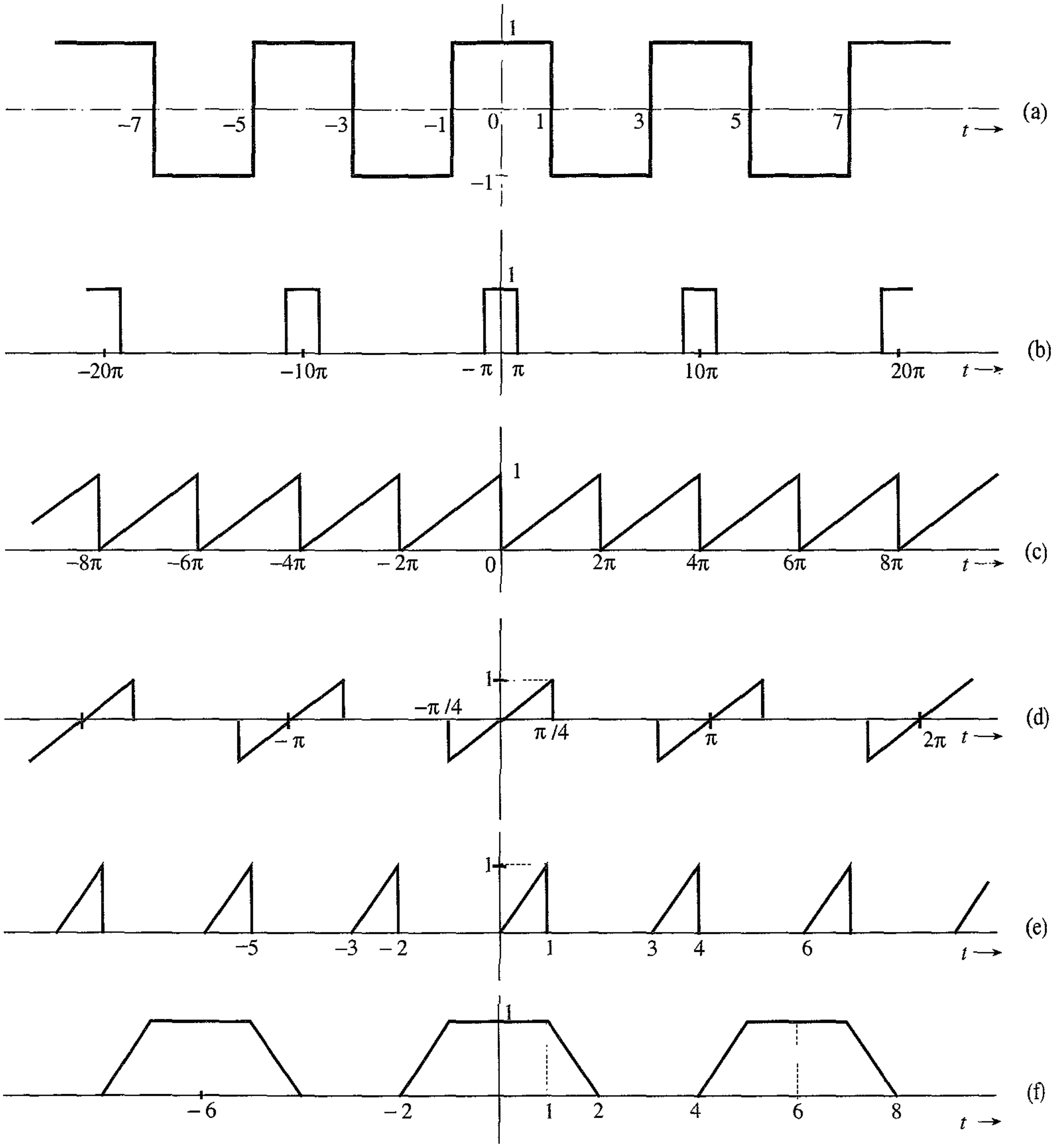

1
-7
-5
-3
-1
0
1
3
5
7
-1
t ⟶
(a)
1
-20π
-10π
- π
π
10π
20π
t ⟶
(b)
1
-8π
-6π
-4π
- 2π
0
2π
4π
6π
8π
t ⟶
(c)
1
-π /4
π /4
- π
π
2π
t ⟶
(d)
1
-5
-3
- 2
1
3
4
6
t ⟶
(e)
1
- 6
- 2
1
2
4
6
8
t ⟶
(f)

Figure P2.8-4

2.8-3 If a periodic signal satisfies certain symmetry conditions, the evaluation of the Fourier series components is somewhat simplified. Show that:

(a) If $g(t) = g(-t)$ (even symmetry), then all the sine terms in the Fourier series vanish ($b_n = 0$).

(b) If $g(t) = -g(-t)$ (odd symmetry), then the dc and all the cosine terms in the Fourier series vanish ($a_0 = a_n = 0$).

Further, show that in each case the Fourier coefficients can be evaluated by integrating the periodic signal over the half-cycle only. This is because the entire information of one cycle is implicit in a half-cycle due to symmetry. *Hint*: If $g_e(t)$ and $g_o(t)$ are even and odd functions, respectively, of t, then (assuming no impulse or its derivative at the origin)

$$\int_{-a}^{a} g_e(t)\,dt = 2\int_{0}^{a} g_e(t)\,dt \qquad \text{and} \qquad \int_{-a}^{a} g_o(t)\,dt = 0$$

Also the product of an even and an odd function is an odd function, the product of two odd functions is an even function, and the product of two even functions is an even function.

2.8-4 For each of the periodic signals shown in Fig. P2.8-4, find the compact trigonometric Fourier series and sketch the amplitude and phase spectra. If either the sine or the cosine terms are absent in the Fourier series, explain why.

2.8-5 (a) Show that an arbitrary function $g(t)$ can be expressed as a sum of an even function $g_e(t)$ and an odd function $g_o(t)$:

$$g(t) = g_e(t) + g_o(t)$$

Hint:

$$g(t) = \underbrace{\frac{1}{2}[g(t) + g(-t)]}_{g_e(t)} + \underbrace{\frac{1}{2}[g(t) - g(-t)]}_{g_o(t)}$$

(b) Determine the odd and even components of the functions: (i) $u(t)$; (ii) $e^{-at}u(t)$; (iii) e^{jt}.

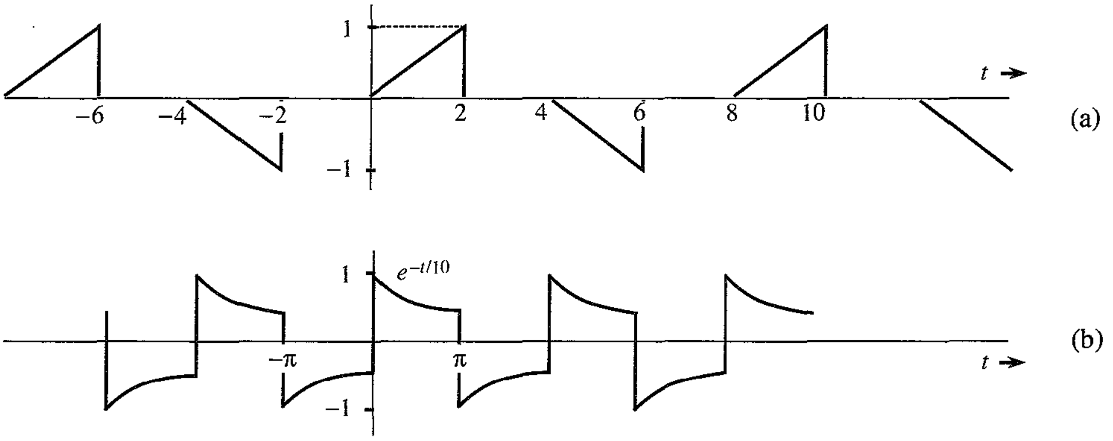


Figure P2.8-6

2.8-6 If the two halves of one period of a periodic signal are of identical shape except that one is the negative of the other, the periodic signal is said to have a **half-wave symmetry**. If a periodic signal $g(t)$ with a period T_0 satisfies the half-wave symmetry condition, then

$$g\left(t - \frac{T_0}{2}\right) = -g(t)$$

In this case, show that all the even-numbered harmonics vanish, and that the odd-numbered harmonic coefficients are given by

$$a_n = \frac{4}{T_0} \int_0^{T_0/2} g(t) \cos n\omega_0 t \, dt \quad \text{and} \quad b_n = \frac{4}{T_0} \int_0^{T_0/2} g(t) \sin n\omega_0 t \, dt$$

Using these results, find the Fourier series for the periodic signals in Fig. P2.8-6.

2.9-1 For each of the periodic signals in Fig. P2.8-4, find exponential Fourier series and sketch the corresponding spectra.

2.9-2 A periodic signal $g(t)$ is expressed by the following Fourier series:

$$g(t) = 3 \cos t + \cos\left(5t - \frac{2\pi}{3}\right) + 2 \cos\left(8t + \frac{2\pi}{3}\right)$$

(a) Sketch the amplitude and phase spectra for the trigonometric series.

(b) By inspection of spectra in part **(a)**, sketch the exponential Fourier series spectra.

(c) By inspection of spectra in part **(b)**, write the exponential Fourier series for $g(t)$.

2.9-3 Figure P2.9-3 shows the trigonometric Fourier spectra of a periodic signal $g(t)$.

(a) By inspection of Fig. P2.9-3, find the trigonometric Fourier series representing $g(t)$.

(b) By inspection of Fig. P2.9-3, sketch the exponential Fourier spectra of $g(t)$.

(c) By inspection of the exponential Fourier spectra obtained in part **(b)**, find the exponential Fourier series for $g(t)$.

(d) Show that the series found in parts **(a)** and **(c)** are equivalent.

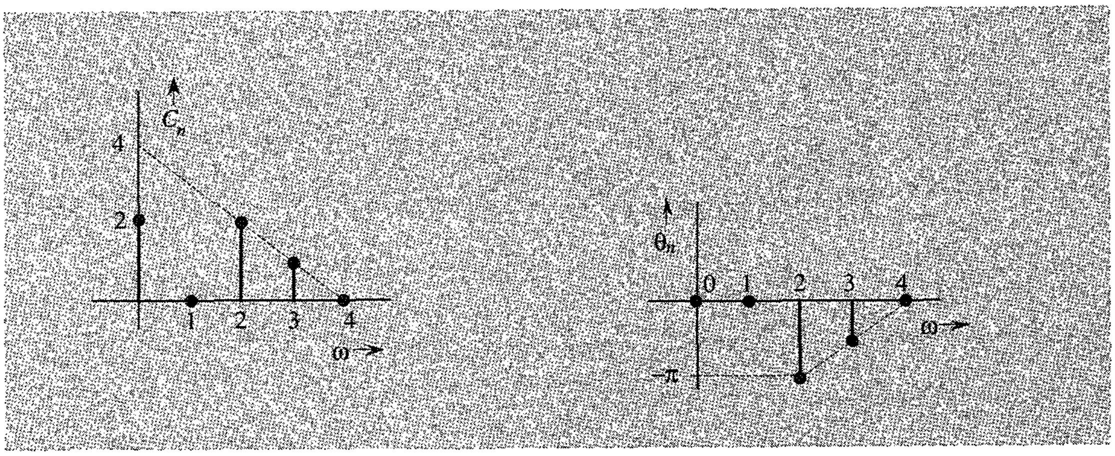


Figure P2.9-3

2.9-4 Show that the coefficients of the exponential Fourier series of an even periodic signal are real and those of an odd periodic signal are imaginary.

3 ANALYSIS AND TRANSMISSION OF SIGNALS

Electrical engineers instinctively think of signals in terms of their frequency spectra and think of systems in terms of their frequency responses. Even teenagers know about audio signals having a bandwidth of 20 kHz and good-quality loud speakers responding up to 20 kHz. This is basically thinking in the frequency domain. In the last chapter we discussed spectral representation of periodic signals (Fourier series). In this chapter we extend this spectral representation to aperiodic signals.

3.1 APERIODIC SIGNAL REPRESENTATION BY FOURIER INTEGRAL

Applying a limiting process, we now show that an aperiodic signal can be expressed as a continuous sum (integral) of everlasting exponentials. To represent an aperiodic signal $g(t)$, such as the one shown in Fig. 3.1a by everlasting exponential signals, let us construct a new periodic signal $g_{T_0}(t)$ formed by repeating the signal $g(t)$ every T_0 seconds, as shown in Fig. 3.1b. The period T_0 is made long enough to avoid overlap between the repeating pulses. The periodic signal $g_{T_0}(t)$ can be represented by an exponential Fourier series. If we let $T_0 \to \infty$, the pulses in the periodic signal repeat after an infinite interval, and therefore

$$\lim_{T_0 \to \infty} g_{T_0}(t) = g(t)$$

Thus, the Fourier series representing $g_{T_0}(t)$ will also represent $g(t)$ in the limit $T_0 \to \infty$. The exponential Fourier series for $g_{T_0}(t)$ is given by

$$g_{T_0}(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \tag{3.1}$$

in which

$$D_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} g_{T_0}(t) e^{-jn\omega_0 t}\, dt \tag{3.2a}$$

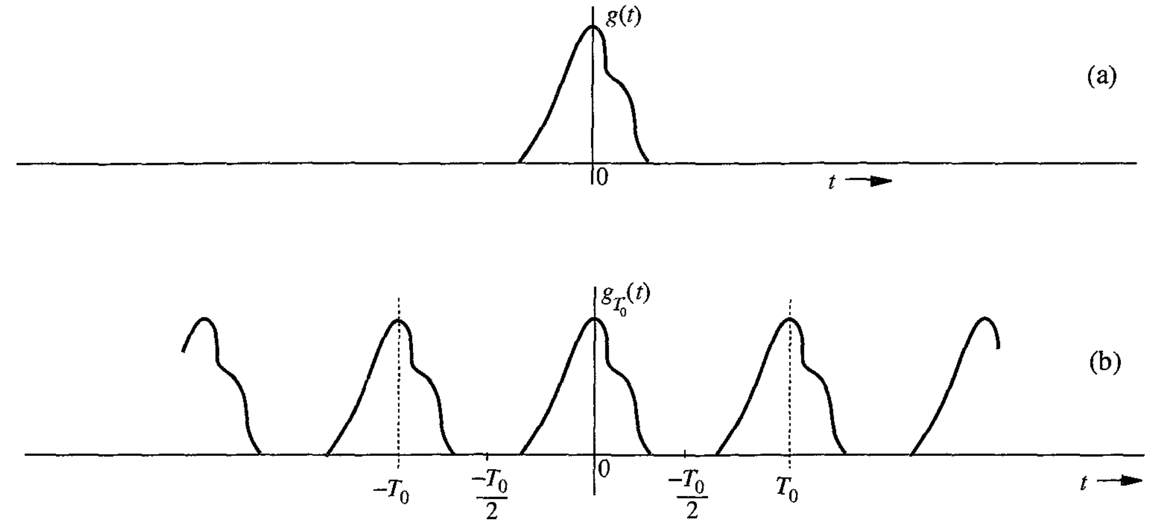


Figure 3.1 Construction of a periodic signal by periodic extension of $g(t)$.

and

$$\omega_0 = \frac{2\pi}{T_0} \tag{3.2b}$$

Observe that integrating $g_{T_0}(t)$ over $(-T_0/2, T_0/2)$ is the same as integrating $g(t)$ over $(-\infty, \infty)$. Therefore, Eq. (3.2a) can be expressed as

$$D_n = \frac{1}{T_0} \int_{-\infty}^{\infty} g(t) e^{-jn\omega_0 t}\, dt \tag{3.2c}$$

It is interesting to see how the nature of the spectrum changes as T_0 increases. To understand this fascinating behavior, let us define $G(\omega)$, a continuous function of ω, as

$$G(\omega) = \int_{-\infty}^{\infty} g(t) e^{-j\omega t}\, dt \tag{3.3}$$

A glance at Eqs. (3.2c) and (3.3) shows that

$$D_n = \frac{1}{T_0} G(n\omega_0) \tag{3.4}$$

This shows that the Fourier coefficients D_n are ($1/T_0$ times) the samples of $G(\omega)$ uniformly spaced at intervals of ω_0 rad/s, as shown in Fig. 3.2a*. Therefore, $(1/T_0)G(\omega)$ is the envelope for the coefficients D_n. We now let $T_0 \to \infty$ by doubling T_0 repeatedly. Doubling T_0 halves the fundamental frequency ω_0, so that there are now twice as many components (samples) in the spectrum. However, by doubling T_0, the envelope $(1/T_0)G(\omega)$ is halved, as shown in Fig. 3.2b. If we continue this process of doubling T_0 repeatedly, the spectrum progressively becomes denser while its magnitude becomes smaller. Note, however, that the relative shape of the envelope remains the same [proportional to $G(\omega)$ in Eq. (3.3)]. In the limit as $T_0 \to \infty$, $\omega_0 \to 0$ and $D_n \to 0$. This means that the spectrum is so dense that the spectral components are spaced at zero (infinitesimal) interval. At the same time, the amplitude of each component is zero (infinitesimal). We have *nothing of everything, yet we have something!* This sounds like *Alice in Wonderland*, but as we shall see, these are the classic characteristics of a very familiar phenomenon.*

---

* For the sake of simplicity we assume D_n and therefore $G(\omega)$ in Fig. 3.2 to be real. The argument, however, is also valid for complex D_n [or $G(\omega)$].

Substitution of Eq. (3.4) in Eq. (3.1) yields

$$g_{T_0}(t) = \sum_{n=-\infty}^{\infty} \frac{G(n\omega_0)}{T_0} e^{jn\omega_0 t} \tag{3.5}$$

As $T_0 \to \infty$, ω_0 becomes infinitesimal ($\omega_0 \to 0$). Because of this, we shall replace ω_0 by a more appropriate notation, $\Delta\omega$. In terms of this new notation, Eq. (3.2b) becomes

$$\Delta\omega = \frac{2\pi}{T_0}$$

and Eq. (3.5) becomes

$$g_{T_0}(t) = \sum_{n=-\infty}^{\infty} \left[\frac{G(n\Delta\omega)\Delta\omega}{2\pi}\right] e^{(jn\Delta\omega)t} \tag{3.6a}$$

Equation (3.6a) shows that $g_{T_0}(t)$ can be expressed as a sum of everlasting exponentials of frequencies $0, \pm\Delta\omega, \pm 2\Delta\omega, \pm 3\Delta\omega, \cdots$ (the Fourier series). The amount of the component of frequency $n\Delta\omega$ is $[G(n\Delta\omega)\Delta\omega]/2\pi$. In the limit as $T_0 \to \infty$, $\Delta\omega \to 0$ and $g_{T_0}(t) \to g(t)$. Therefore,

$$g(t) = \lim_{T_0\to\infty} g_{T_0}(t) = \lim_{\Delta\omega\to 0} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} G(n\Delta\omega) e^{(jn\Delta\omega)t} \Delta\omega \tag{3.6b}$$

The sum on the right-hand side of Eq. (3.6b) can be viewed as the area under the function $G(\omega)e^{j\omega t}$, as shown in Fig. 3.3. Therefore,

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) e^{j\omega t} d\omega \tag{3.7}$$

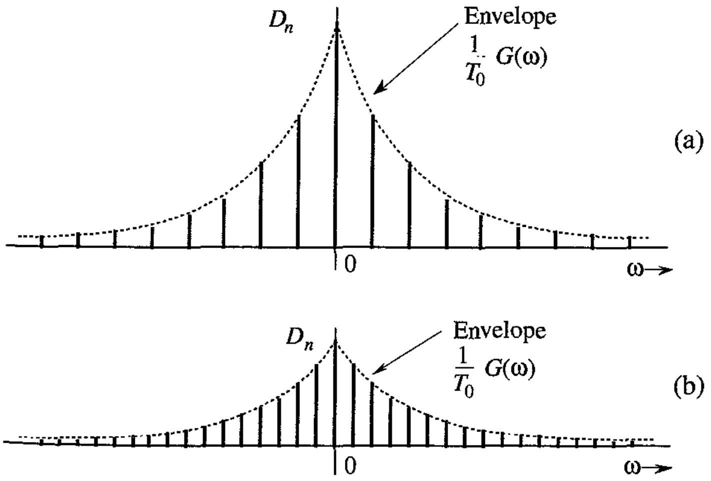


Figure 3.2 Change in the Fourier spectrum when the period T_0 in Fig. 3.1 is doubled.

---

* You may consider this as an irrefutable proof of the proposition that 0% ownership of everything is better than 100% ownership of nothing!

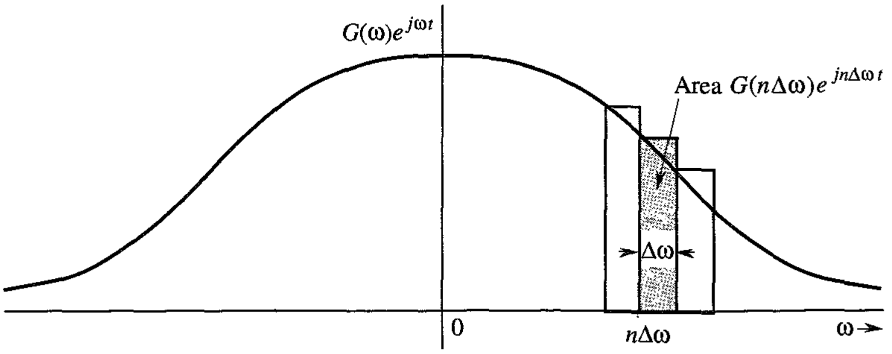


Figure 3.3 The Fourier series becomes the Fourier integral in the limit as $T_0 \rightarrow \infty$.

The integral on the right-hand side is called the **Fourier integral**. We have now succeeded in representing an aperiodic signal $g(t)$ by a Fourier integral* (rather than a Fourier series). This integral is basically a Fourier series (in the limit) with fundamental frequency $\Delta\omega \rightarrow 0$, as seen from Eq. (3.6). The amount of the exponential $e^{jn\Delta\omega t}$ is $G(n\Delta\omega)\Delta\omega/2\pi$. Thus, the function $G(\omega)$ given by Eq. (3.3) acts as a spectral function.

We call $G(\omega)$ the **direct** Fourier transform of $g(t)$, and $g(t)$ the **inverse** Fourier transform of $G(\omega)$. The same information is conveyed by the statement that $g(t)$ and $G(\omega)$ are a Fourier transform pair. Symbolically, this is expressed as

$$G(\omega) = \mathcal{F}[g(t)] \qquad \text{and} \qquad g(t) = \mathcal{F}^{-1}[G(\omega)]$$

or

$$g(t) \Longleftrightarrow G(\omega)$$

To recapitulate,

$$G(\omega) = \int_{-\infty}^{\infty} g(t)e^{-j\omega t}\, dt \tag{3.8a}$$

and

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)e^{j\omega t}\, d\omega \tag{3.8b}$$

It is helpful to keep in mind that the Fourier integral in Eq. (3.8b) is of the nature of a Fourier series with fundamental frequency $\Delta\omega$ approaching zero [Eq. (3.6b)]. Therefore, most of the discussion and properties of Fourier series apply to the Fourier transform as well. We can plot the spectrum $G(\omega)$ as a function of ω. Since $G(\omega)$ is complex, we have both amplitude and angle (or phase) spectra:

$$G(\omega) = |G(\omega)|e^{j\theta_g(\omega)}$$

in which $|G(\omega)|$ is the amplitude and $\theta_g(\omega)$ is the angle (or phase) of $G(\omega)$. From Eq. (3.8a),

$$G(-\omega) = \int_{-\infty}^{\infty} g(t)e^{j\omega t}\, dt$$

---

* This should not be considered as a rigorous proof of Eq. (3.7). The situation is not as simple as we have made it appear.[1]

Conjugate Symmetry Property

From this equation and Eq. (3.8a), it follows that if $g(t)$ is a real function of t, then $G(\omega)$ and $G(-\omega)$ are complex conjugates, that is,

$$G(-\omega) = G^*(\omega) \tag{3.9}$$

Therefore,

$$|G(-\omega)| = |G(\omega)| \tag{3.10a}$$

$$\theta_g(-\omega) = -\theta_g(\omega) \tag{3.10b}$$

Thus, for real $g(t)$, the amplitude spectrum $|G(\omega)|$ is an even function, and the phase spectrum $\theta_g(\omega)$ is an odd function of ω. This property (the **conjugate symmetry property**) is valid only for real $g(t)$. These results were derived earlier for the Fourier spectrum of a periodic signal [Eqs. (2.85)] and should come as no surprise. *The transform $G(\omega)$ is the frequency-domain specification of $g(t)$.*

EXAMPLE 3.1 Find the Fourier transform of $e^{-at}u(t)$.

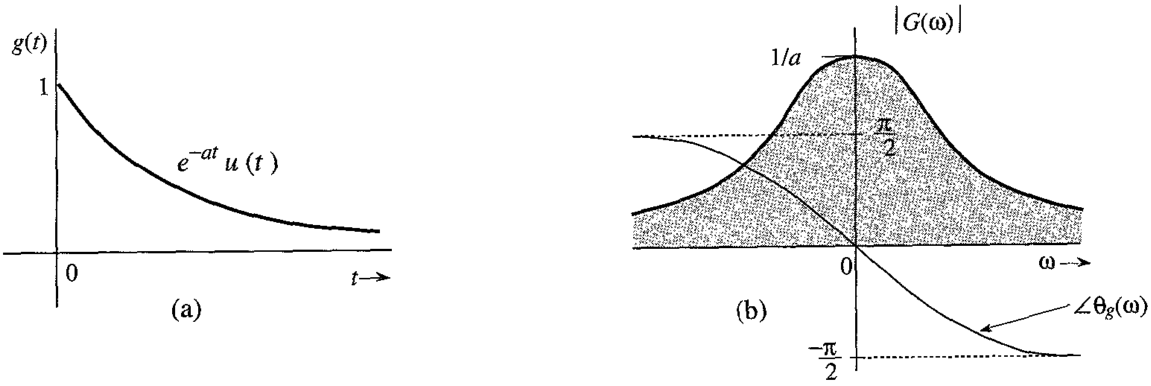


Figure 3.4 $e^{-at}u(t)$ and its Fourier spectra.

By definition [Eq. (3.8a)],

$$G(\omega) = \int_{-\infty}^{\infty} e^{-at}u(t)e^{-j\omega t}\,dt = \int_{0}^{\infty} e^{-(a+j\omega)t}\,dt = \frac{-1}{a+j\omega}e^{-(a+j\omega)t}\Big|_0^{\infty}$$

But $|e^{-j\omega t}| = 1$. Therefore, as $t \rightarrow \infty$, $e^{-(a+j\omega)t} = e^{-at}e^{-j\omega t} = 0$ if $a > 0$. Therefore,

$$G(\omega) = \frac{1}{a+j\omega} \qquad a > 0 \tag{3.11a}$$

Expressing $a + j\omega$ in the polar form as $\sqrt{a^2+\omega^2}\,e^{j\tan^{-1}(\frac{\omega}{a})}$, we obtain

$$G(\omega) = \frac{1}{\sqrt{a^2+\omega^2}}e^{-j\tan^{-1}(\frac{\omega}{a})} \tag{3.11b}$$

Therefore,

$$|G(\omega)| = \frac{1}{\sqrt{a^2+\omega^2}} \qquad \text{and} \qquad \theta_g(\omega) = -\tan^{-1}\left(\frac{\omega}{a}\right)$$

The amplitude spectrum $|G(\omega)|$ and the phase spectrum $\theta_g(\omega)$ are shown in Fig. 3.4b. Observe that $|G(\omega)|$ is an even function of ω, and $\theta_g(\omega)$ is an odd function of ω, as expected.

Existence of the Fourier Transform

In Example 3.1 we observed that when $a < 0$, the Fourier integral for $e^{-at}u(t)$ does not converge. Hence, the Fourier transform for $e^{-at}u(t)$ does not exist if $a < 0$ (growing exponential). Clearly, not all signals are Fourier transformable. The existence of the Fourier transform is assured for any $g(t)$ satisfying the Dirichlet conditions mentioned in Sec. 2.8. The first of these conditions is*

$$\int_{-\infty}^{\infty} |g(t)|\, dt < \infty \tag{3.12}$$

To show this, recall that $|e^{-j\omega t}| = 1$. Hence, from Eq. (3.8a) we obtain

$$|G(\omega)| \leq \int_{-\infty}^{\infty} |g(t)|\, dt$$

This shows that the existence of the Fourier transform is assured if condition (3.12) is satisfied. Otherwise, there is no guarantee. We have seen in Example 3.1 that for an exponentially growing signal (which violates this condition) the Fourier transform does not exist. Although this condition is sufficient, it is not necessary for the existence of the Fourier transform of a signal. For example, the signal $(\sin at)/t$, violates condition (3.12), but does have a Fourier transform. Any signal that can be generated in practice satisfies the Dirichlet conditions and therefore has a Fourier transform. Thus, the physical existence of a signal is a sufficient condition for the existence of its transform.

Linearity of the Fourier Transform

The Fourier transform is linear; that is, if

$$g_1(t) \Longleftrightarrow G_1(\omega) \qquad \text{and} \qquad g_2(t) \Longleftrightarrow G_2(\omega)$$

then

$$a_1 g_1(t) + a_2 g_2(t) \Longleftrightarrow a_1 G_1(\omega) + a_2 G_2(\omega) \tag{3.13}$$

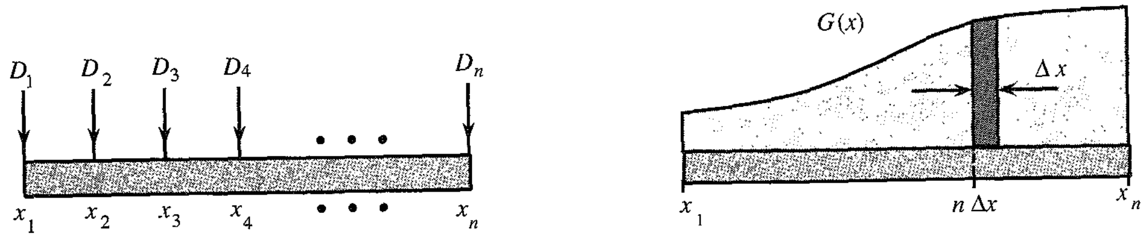
The proof is trivial and follows directly from Eq. (3.8a). This result can be extended to any finite number of terms.

3.1.1 Physical Appreciation of the Fourier Transform

In understanding any aspect of the Fourier transform, we should remember that Fourier representation is a way of expressing a signal in terms of everlasting sinusoids, or exponentials.

---

* The remaining Dirichlet conditions are as follows: In any finite interval, $g(t)$ may have only a finite number of maxima and minima and a finite number of finite discontinuities. When these conditions are satisfied, the Fourier integral on the right-hand side of Eq. (3.8b) converges to $g(t)$ at all points where $g(t)$ is continuous and converges to the average of the right-hand and left-hand limits of $g(t)$ at points where $g(t)$ is discontinuous.



(a) (b)

Figure 3.5 Analogy for Fourier transform.

The Fourier spectrum of a signal indicates the relative amplitudes and phases of the sinusoids that are required to synthesize that signal. A periodic signal Fourier spectrum has finite amplitudes and exists at discrete frequencies (ω_0 and its multiples). Such a spectrum is easy to visualize, but the spectrum of an aperiodic signal is not easy to visualize because it has a continuous spectrum that exists at every frequency. The continuous spectrum concept can be appreciated by considering an analogous, more tangible phenomenon. One familiar example of a continuous distribution is the loading of a beam. Consider a beam loaded with weights $D_1, D_2, D_3, \cdots, D_n$ units at the uniformly spaced points $x_1, x_2, \cdots, x_n$, as shown in Fig. 3.5a. The total load W_T on the beam is given by the sum of these loads at each of the n points:

$$W_T = \sum_{i=1}^{n} D_i$$

Consider now the case of a continuously loaded beam, as shown in Fig. 3.5b. In this case, although there appears to be a load at every point, the load at any one point is zero. This does not mean that there is no load on the beam. A meaningful measure of load in this situation is not the load at a point, but rather the loading density per unit length at that point. Let $G(x)$ be the loading density per unit length of beam. This means that the load over a beam length Δx $(\Delta x \to 0)$ at some point x is $G(x)\Delta x$. To find the total load on the beam, we divide the beam into segments of interval Δx $(\Delta x \to 0)$. The load over the nth such segment of length Δx is $[G(n\Delta x)]\,\Delta x$. The total load W_T is given by

$$\begin{aligned} W_T &= \lim_{\Delta x \to 0} \sum_{x_1}^{x_n} G(n\Delta x)\,\Delta x \\ &= \int_{x_1}^{x_n} G(x)\,dx \end{aligned}$$

In the case of discrete loading (Fig. 3.5a), the load exists only at the n discrete points. At other points there is no load. On the other hand, in the continuously loaded case, the load exists at every point, but at any specific point x the load is zero. The load over a small interval Δx, however, is $[G(n\Delta x)]\,\Delta x$ (Fig. 3.5b). Thus, even though the load at a point x is zero, the relative load at that point is $G(x)$.

An exactly analogous situation exists in the case of a signal spectrum. When $g(t)$ is periodic, the spectrum is discrete, and $g(t)$ can be expressed as a sum of discrete exponentials with finite amplitudes:

$$g(t) = \sum_{n} D_n e^{jn\omega_0 t}$$

For an aperiodic signal, the spectrum becomes continuous; that is, the spectrum exists for every value of ω, but the amplitude of each component in the spectrum is zero. The meaningful measure here is not the amplitude of a component of some frequency but the spectral density per unit bandwidth. From Eq. (3.6b) it is clear that $g(t)$ is synthesized by adding exponentials of the form $e^{jn\Delta\omega t}$, in which the contribution by any one exponential component is zero. But the contribution by exponentials in an infinitesimal band $\Delta\omega$ located at $\omega = n\Delta\omega$ is $(1/2\pi)G(n\Delta\omega)\Delta\omega$, and the addition of all these components yields $g(t)$ in the integral form:

$$g(t) = \lim_{\Delta\omega \to 0} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} G(n\Delta\omega)e^{(jn\Delta\omega)t}\Delta\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)e^{j\omega t}\,d\omega$$

The contribution by components within the band $d\omega$ is $(1/2\pi)G(\omega)\,d\omega = G(\omega)\,df$, in which df is the bandwidth in hertz. Clearly $G(\omega)$ is the **spectral density** per unit bandwidth (in hertz). This also means that even if the amplitude of any one component is zero, the relative amount of a component of frequency ω is $G(\omega)$. Although $G(\omega)$ is a spectral density, in practice it is customarily called the **spectrum** of $g(t)$ rather than the spectral density of $g(t)$. Deferring to this convention, we shall call $G(\omega)$ the Fourier spectrum (or Fourier transform) of $g(t)$.

A Marvelous Balancing Act

An important point to remember here is that $g(t)$ is represented (or synthesized) by exponentials or sinusoids that are everlasting (not causal). This leads to a rather fascinating picture when we try to visualize the synthesis of a time-limited pulse signal $g(t)$ (Fig. 3.6) according to the sinusoidal components in its Fourier spectrum. The signal $g(t)$ exists only over an interval $(a,\ b)$ and is zero outside this interval. The spectrum of $g(t)$ contains an infinite number of exponentials (or sinusoids) which start at $t = -\infty$ and continue forever. The amplitudes and phases of these components are such that they add up exactly to $g(t)$ over the finite interval $(a,\ b)$ and add up to zero everywhere outside this interval. Juggling with such a perfect and delicate balance of amplitudes and phases of an infinite number of components boggles the human imagination. Yet, the Fourier transform accomplishes it routinely, without much thinking on our part. Indeed, we become so involved in mathematical manipulations that we fail to notice this marvel.

3.2 TRANSFORMS OF SOME USEFUL FUNCTIONS

For convenience, we now introduce a compact notation for some useful functions such as gate, triangle, and interpolation functions.

Unit Gate Function

We define a unit gate function rect (x) as a gate pulse of unit height and unit width, centered at the origin, as shown in Fig. 3.7a:

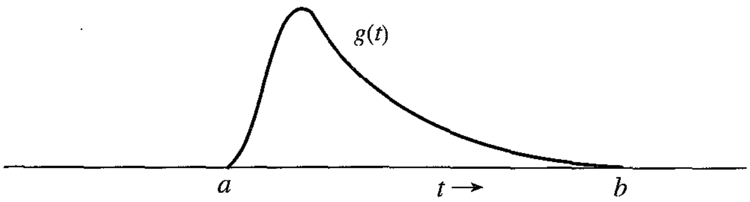


Figure 3.6 A time-limited pulse.

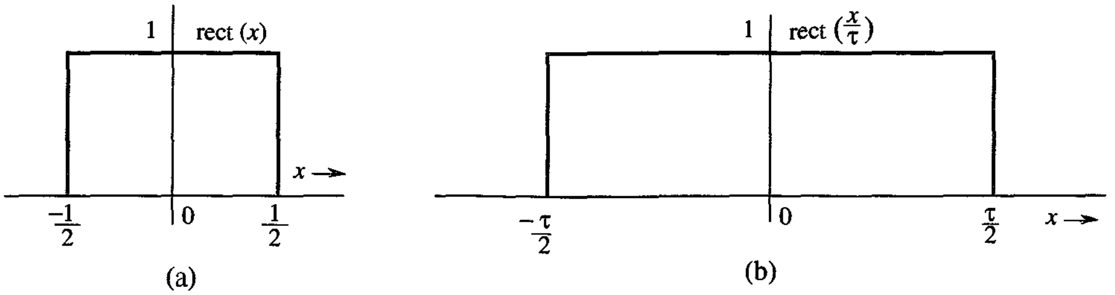

1
rect (x)
$x \rightarrow$
$-\frac{1}{2}$
0
$\frac{1}{2}$
(a)
1
rect $\left(\frac{x}{\tau}\right)$
$-\frac{\tau}{2}$
0
$\frac{\tau}{2}$
$x \rightarrow$
(b)

Figure 3.7 Gate pulse.

$$\operatorname{rect}(x) = \begin{cases} 0 & |x| > \frac{1}{2} \\ \frac{1}{2} & |x| = \frac{1}{2} \\ 1 & |x| < \frac{1}{2} \end{cases} \tag{3.14}$$

The gate pulse in Fig. 3.7b is the unit gate pulse rect (x) expanded by a factor τ and therefore can be expressed as rect (x/τ) (see Sec. 2.3.2). Observe that τ, the denominator of the argument of rect (x/τ), indicates the width of the pulse.

Unit Triangle Function

We define a unit triangle function $\Delta(x)$ as a triangular pulse of unit height and unit width, centered at the origin, as shown in Fig. 3.8a:

$$\Delta(x) = \begin{cases} 0 & |x| > \frac{1}{2} \\ 1 - 2|x| & |x| < \frac{1}{2} \end{cases} \tag{3.15}$$

The pulse in Fig. 3.8b is $\Delta(x/\tau)$. Observe that here, as for the gate pulse, the denominator τ of the argument of $\Delta(x/\tau)$ indicates the pulse width.

Interpolation Function sinc (x)

The function sin x/x is the "sine over argument" function denoted by sinc (x).* This function

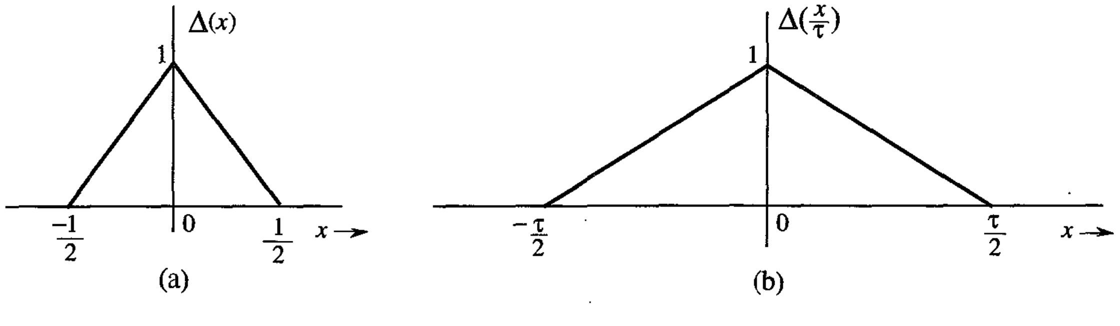

$\Delta(x)$
1
$-\frac{1}{2}$
0
$\frac{1}{2}$
$x \rightarrow$
(a)
$\Delta\left(\frac{x}{\tau}\right)$
1
$-\frac{\tau}{2}$
0
$\frac{\tau}{2}$
$x \rightarrow$
(b)

Figure 3.8 Triangle pulse.

---

* sinc (x) is also denoted by Sa (x) in the literature. Some authors define sinc (x) as

$$\operatorname{sinc}(x) = \frac{\sin \pi x}{\pi x}$$

plays an important role in signal processing. It is also known as the **filtering** or **interpolating function**. We define

$$\operatorname{sinc}(x) = \frac{\sin x}{x} \tag{3.16}$$

Inspection of Eq. (3.16) shows that

1. sinc (x) is an even function of x.
2. sinc $(x) = 0$ when $\sin x = 0$ except at $x = 0$, where it is indeterminate. This means that sinc $(x) = 0$ for $x = \pm\pi, \pm 2\pi, \pm 3\pi, \cdots$.
3. Using L'Hôpital's rule, we find sinc $(0) = 1$.
4. sinc (x) is the product of an oscillating signal $\sin x$ (of period 2π) and a monotonically decreasing function $1/x$. Therefore, sinc (x) exhibits sinusoidal oscillations of period 2π, with amplitude decreasing continuously as $1/x$.

Figure 3.9a shows sinc (x). Observe that sinc $(x) = 0$ for values of x that are positive and negative integral multiples of π. Figure 3.9b shows sinc $(3\omega/7)$. The argument $3\omega/7 = \pi$ when $\omega = 7\pi/3$. Therefore, the first zero of this function occurs at $\omega = 7\pi/3$.

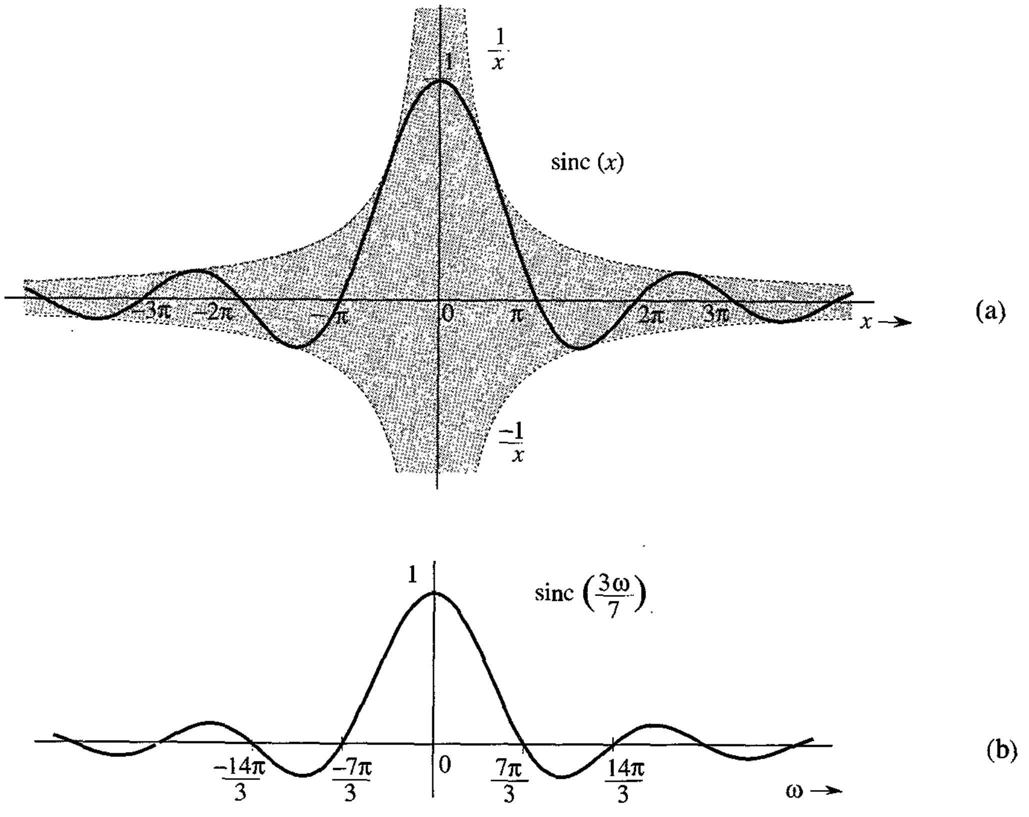


Figure 3.9 Sinc pulse.

EXAMPLE 3.2 Find the Fourier transform of $g(t) = \operatorname{rect}(t/\tau)$ (Fig. 3.10a).

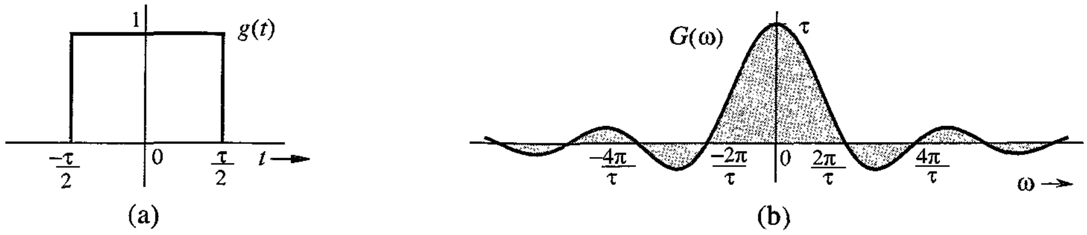


Figure 3.10 Gate pulse and its Fourier spectrum.

We have

$$G(\omega) = \int_{-\infty}^{\infty} \text{rect}\left(\frac{t}{\tau}\right) e^{-j\omega t}\,dt$$

Since rect $(t/\tau) = 1$ for $|t| < \tau/2$, and since it is zero for $|t| > \tau/2$,

$$\begin{aligned} G(\omega) &= \int_{-\tau/2}^{\tau/2} e^{-j\omega t}\,dt \\ &= -\frac{1}{j\omega}\left(e^{-j\omega\tau/2} - e^{j\omega\tau/2}\right) = \frac{2\sin(\omega\tau/2)}{\omega} \\ &= \tau\frac{\sin(\omega\tau/2)}{(\omega\tau/2)} = \tau \operatorname{sinc}\left(\frac{\omega\tau}{2}\right) \end{aligned}$$

Therefore,

$$\text{rect}\left(\frac{t}{\tau}\right) \Longleftrightarrow \tau \operatorname{sinc}\left(\frac{\omega\tau}{2}\right) \tag{3.17}$$

Recall that sinc $(x) = 0$ when $x = \pm n\pi$. Hence, sinc $(\omega\tau/2) = 0$ when $\omega\tau/2 = \pm n\pi$; that is, when $\omega = \pm 2n\pi/\tau$ $(n = 1, 2, 3, \cdots)$, as shown in Fig. 3.10b. Observe that in this case $G(\omega)$ happens to be real. Hence, we may convey the spectral information by a single plot of $G(\omega)$ shown in Fig. 3.10b.

Bandwidth of rect $(\frac{t}{\tau})$

The spectrum $G(\omega)$ in Fig. 3.10 peaks at $\omega = 0$ and decays at higher frequencies. Therefore, rect (t/τ) is a low-pass signal with most of the signal energy in lower frequency components. **Signal bandwidth** is the difference between the highest (significant) frequency and the lowest (significant) frequency in the signal spectrum. Strictly speaking, because the spectrum extends from 0 to ∞, the bandwidth is ∞ in the present case. However, much of the spectrum is concentrated within the first lobe (from $\omega = 0$ to $\omega = 2\pi/\tau$), and we may consider $\omega = 2\pi/\tau$ to be the highest (significant) frequency in the spectrum. Therefore, a rough estimate of the bandwidth* of a rectangular pulse of width τ seconds is $2\pi/\tau$ rad/s, or $1/\tau$ Hz. Note the reciprocal relationship of the pulse width with its bandwidth. We shall observe later that this result is true in general.

---

* To compute the bandwidth, we must consider the spectrum only for positive values of ω. The trigonometric spectrum exists only for positive frequencies. The negative frequencies occur because we use exponential spectra for mathematical convenience. Each sinusoid $\cos\omega_n t$ appears as a sum of two exponential components $e^{j\omega_n t}$ and $e^{-j\omega_n t}$ with frequencies ω_n and $-\omega_n$, respectively. But in reality, there is only one component of frequency ω_n.

EXAMPLE 3.3 Find the Fourier transform of the unit impulse $\delta(t)$.

Using the sampling property of the impulse [Eq. (2.19a)], we obtain

$$\mathcal{F}[\delta(t)] = \int_{-\infty}^{\infty} \delta(t)e^{-j\omega t}dt = 1 \tag{3.18a}$$

or

$$\delta(t) \Longleftrightarrow 1 \tag{3.18b}$$

Figure 3.11 shows $\delta(t)$ and its spectrum.

g(t) = δ(t)
0
t→
(a)
G(ω) = 1
1
0
ω→
(b)

Figure 3.11 Unit impulse and its Fourier spectrum.

EXAMPLE 3.4 Find the inverse Fourier transform of $\delta(\omega)$.

From Eq. (3.8b) and the sampling property of the impulse function,

$$\mathcal{F}^{-1}[\delta(\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta(\omega)e^{j\omega t}\, d\omega = \frac{1}{2\pi}$$

Therefore,

$$\frac{1}{2\pi} \Longleftrightarrow \delta(\omega) \tag{3.19a}$$

or

$$1 \Longleftrightarrow 2\pi\delta(\omega) \tag{3.19b}$$

This shows that the spectrum of a constant signal $g(t) = 1$ is an impulse $2\pi\delta(\omega)$, as shown in Fig. 3.12.

The result [Eq. (3.19b)] also could have been anticipated on qualitative grounds. Recall that the Fourier transform of $g(t)$ is a spectral representation of $g(t)$ in terms of everlasting exponential components of the form $e^{j\omega t}$. Now to represent a constant signal $g(t) = 1$, we need a single everlasting exponential[*] $e^{j\omega t}$ with $\omega = 0$. This results in a spectrum at a single frequency $\omega = 0$. Another way of looking at the situation is that $g(t) = 1$ is a dc signal which has a single frequency $\omega = 0$ (dc).

---

[*] The constant multiplier 2π in the spectrum $[G(\omega) = 2\pi\delta(\omega)]$ may be a bit puzzling. Since $1 = e^{j\omega t}$ with $\omega = 0$, it appears that the Fourier transform of $g(t) = 1$ should be an impulse of strength unity rather than 2π. Recall, however, that in the Fourier transform $g(t)$ is synthesized not by exponentials of amplitude $G(n\Delta\omega)\Delta\omega$, but of amplitude $1/2\pi$ times $G(n\Delta\omega)\Delta\omega$, as seen from Eq. (3.6b). Had we used variable f (in hertz) instead of ω, the spectrum would have been a unit impulse.

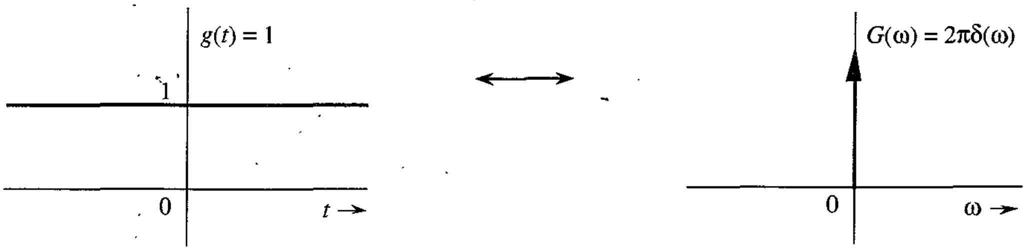


Figure 3.12 Constant (dc) signal and its Fourier spectrum.

If an impulse at $\omega = 0$ is a spectrum of a dc signal, what does an impulse at $\omega = \omega_0$ represent? We shall answer this question in the next example.

EXAMPLE 3.5 Find the inverse Fourier transform of $\delta(\omega - \omega_0)$.

Using the sampling property of the impulse function, we obtain

$$\mathcal{F}^{-1}[\delta(\omega - \omega_0)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta(\omega - \omega_0)e^{j\omega t}\,d\omega = \frac{1}{2\pi}e^{j\omega_0 t}$$

Therefore,

$$\frac{1}{2\pi}e^{j\omega_0 t} \Longleftrightarrow \delta(\omega - \omega_0)$$

or

$$e^{j\omega_0 t} \Longleftrightarrow 2\pi\delta(\omega - \omega_0) \tag{3.20a}$$

This result shows that the spectrum of an everlasting exponential $e^{j\omega_0 t}$ is a single impulse at $\omega = \omega_0$. We reach the same conclusion by qualitative reasoning. To represent the everlasting exponential $e^{j\omega_0 t}$, we need a single everlasting exponential $e^{j\omega t}$ with $\omega = \omega_0$. Therefore, the spectrum consists of a single component at frequency $\omega = \omega_0$.

From Eq. (3.20a) it follows that

$$e^{-j\omega_0 t} \Longleftrightarrow 2\pi\delta(\omega + \omega_0) \tag{3.20b}$$

EXAMPLE 3.6 Find the Fourier transforms of the everlasting sinusoid $\cos \omega_0 t$.

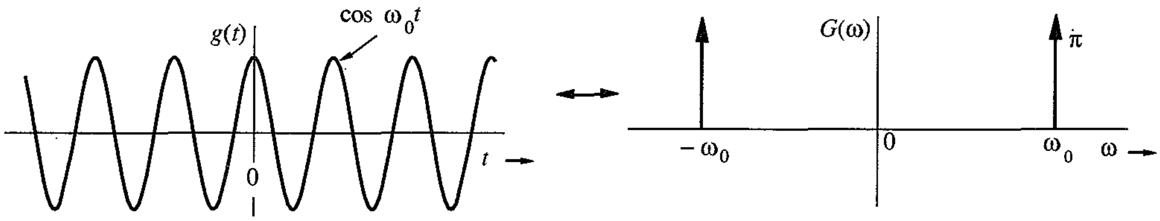


Figure 3.13 Cosine signal and its Fourier spectrum.

Recall the Euler formula

$$\cos \omega_0 t = \frac{1}{2}(e^{j\omega_0 t} + e^{-j\omega_0 t})$$

Adding Eqs. (3.20a) and (3.20b), and using the above formula, we obtain

$$\cos \omega_0 t \Longleftrightarrow \pi[\delta(\omega+\omega_0)+\delta(\omega-\omega_0)] \tag{3.21}$$

The spectrum of cos $\omega_0 t$ consists of two impulses at ω_0 and $-\omega_0$, as shown in Fig. 3.13. The result also follows from qualitative reasoning. An everlasting sinusoid cos $\omega_0 t$ can be synthesized by two everlasting exponentials, $e^{j\omega_0 t}$ and $e^{-j\omega_0 t}$. Therefore, the Fourier spectrum consists of only two components of frequencies ω_0 and $-\omega_0$.

EXAMPLE 3.7 Find the Fourier transform of the sign function sgn t (pronounced signum t), shown in Fig. 3.14. Its value is +1 or −1, depending on whether t is positive or negative:

$$\operatorname{sgn} t = \begin{cases} 1 & t>0 \\ -1 & t<0 \end{cases} \tag{3.22}$$

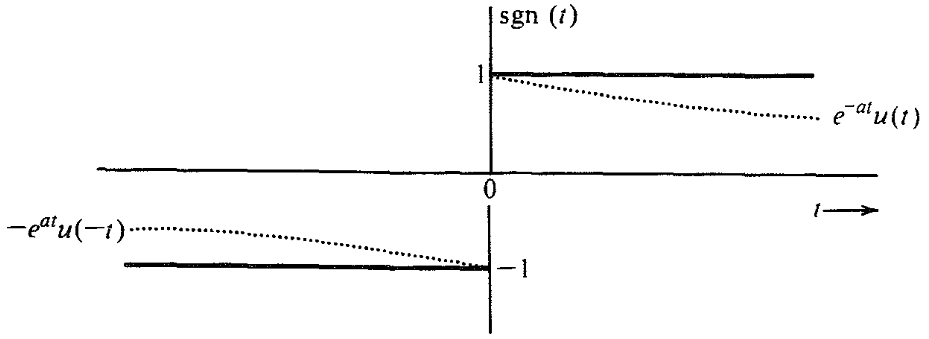


Figure 3.14 Sign function.

The transform of sgn t can be obtained by considering sgn t as a sum of two exponentials, as shown in Fig. 3.14, in the limit as $a \to 0$:

$$\operatorname{sgn} t = \lim_{a\to 0}\left[e^{-at}u(t) - e^{at}u(-t)\right]$$

Therefore,

$$\begin{aligned} \mathcal{F}[\operatorname{sgn} t] &= \lim_{a\to 0}\left\{\mathcal{F}[e^{-at}u(t)] - \mathcal{F}[e^{at}u(-t)]\right\} \\ &= \lim_{a\to 0}\left(\frac{1}{a+j\omega} - \frac{1}{a-j\omega}\right) \qquad \text{(see pairs 1 and 2 in Table 3.1)} \\ &= \lim_{a\to 0}\left(\frac{-2j\omega}{a^2+\omega^2}\right) = \frac{2}{j\omega} \end{aligned} \tag{3.23}$$

3.3 SOME PROPERTIES OF THE FOURIER TRANSFORM

We now study some of the important properties of the Fourier transform and their implications as well as their applications. Before embarking on this study, it is important to point out a pervasive aspect of the Fourier transform—the **time-frequency duality**.

Table 3.1

Short Table of Fourier Transforms

	$g(t)$	$G(\omega)$	
1	$e^{-at}u(t)$	$\dfrac{1}{a+j\omega}$	$a>0$
2	$e^{at}u(-t)$	$\dfrac{1}{a-j\omega}$	$a>0$
3	$e^{-a\|t\|}$	$\dfrac{2a}{a^2+\omega^2}$	$a>0$
4	$te^{-at}u(t)$	$\dfrac{1}{(a+j\omega)^2}$	$a>0$
5	$t^n e^{-at}u(t)$	$\dfrac{n!}{(a+j\omega)^{n+1}}$	$a>0$
6	$\delta(t)$	1	
7	1	$2\pi\delta(\omega)$	
8	$e^{j\omega_0 t}$	$2\pi\delta(\omega-\omega_0)$	
9	$\cos\,\omega_0 t$	$\pi[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]$	
10	$\sin\,\omega_0 t$	$j\pi[\delta(\omega+\omega_0)-\delta(\omega-\omega_0)]$	
11	$u(t)$	$\pi\delta(\omega)+\dfrac{1}{j\omega}$	
12	$\operatorname{sgn} t$	$\dfrac{2}{j\omega}$	
13	$\cos\omega_0 t\; u(t)$	$\dfrac{\pi}{2}[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]+\dfrac{j\omega}{\omega_0^2-\omega^2}$	
14	$\sin\omega_0 t\; u(t)$	$\dfrac{\pi}{2j}[\delta(\omega-\omega_0)-\delta(\omega+\omega_0)]+\dfrac{\omega_0}{\omega_0^2-\omega^2}$	
15	$e^{-at}\sin\omega_0 t\; u(t)$	$\dfrac{\omega_0}{(a+j\omega)^2+\omega_0^2}$	$a>0$
16	$e^{-at}\cos\omega_0 t\; u(t)$	$\dfrac{a+j\omega}{(a+j\omega)^2+\omega_0^2}$	$a>0$
17	$\operatorname{rect}\left(\dfrac{t}{\tau}\right)$	$\tau\operatorname{sinc}\left(\dfrac{\omega\tau}{2}\right)$	
18	$\dfrac{W}{\pi}\operatorname{sinc}(Wt)$	$\operatorname{rect}\left(\dfrac{\omega}{2W}\right)$	
19	$\Delta\left(\dfrac{t}{\tau}\right)$	$\dfrac{\tau}{2}\operatorname{sinc}^2\left(\dfrac{\omega\tau}{4}\right)$	
20	$\dfrac{W}{2\pi}\operatorname{sinc}^2\left(\dfrac{Wt}{2}\right)$	$\Delta\left(\dfrac{\omega}{2W}\right)$	
21	$\sum_{n=-\infty}^{\infty}\delta(t-nT)$	$\omega_0\sum_{n=-\infty}^{\infty}\delta(\omega-n\omega_0)$	$\omega_0=\dfrac{2\pi}{T}$
22	$e^{-t^2/2\sigma^2}$	$\sigma\sqrt{2\pi}e^{-\sigma^2\omega^2/2}$	

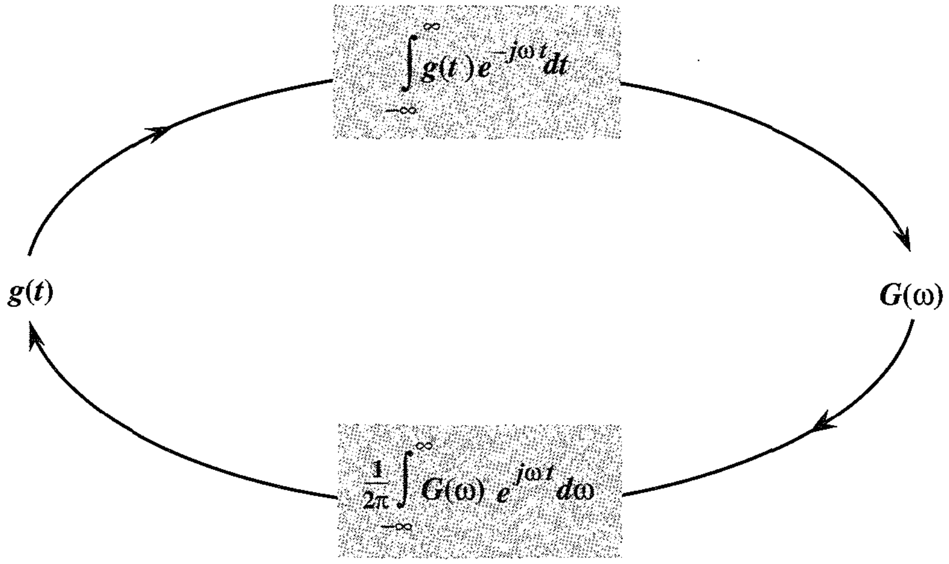

$\int_{-\infty}^{\infty} g(t)e^{-j\omega t}dt$
$g(t)$
$G(\omega)$
$\frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)\, e^{j\omega t} d\omega$

Figure 3.15 Near symmetry between direct and inverse Fourier transforms.

3.3.1 Symmetry of Direct and Inverse Transform Operations—Time-Frequency Duality

Equations (3.8) show an interesting fact: the direct and the inverse transform operations are remarkably similar. These operations, required to go from $g(t)$ to $G(\omega)$ and then from $G(\omega)$ to $g(t)$, are shown graphically in Fig. 3.15. There are only two minor differences in these operations: the factor 2π appears only in the inverse operator, and the exponential indices in the two operations have opposite signs. Otherwise the two operations are symmetrical.* This observation has far-reaching consequences in the study of Fourier transform. It is the basis of the so-called duality of time and frequency. *The duality principle may be compared with a photograph and its negative. A photograph can be obtained from its negative, and by using an identical procedure, the negative can be obtained from the photograph.* For any result or relationship between $g(t)$ and $G(\omega)$, there exists a dual result or relationship, obtained by interchanging the roles of $g(t)$ and $G(\omega)$ in the original result (along with some minor modifications arising because of the factor 2π and a sign change). For example, the time-shifting property, to be proved later, states that if $g(t) \Longleftrightarrow G(\omega)$, then

$$g(t - t_0) \Longleftrightarrow G(\omega)e^{-j\omega t_0}$$

The dual of this property (the frequency-shifting property) states that

---

* Of the two differences, the former can be eliminated by a change of variable from ω to f (in hertz). In this case,

$$\omega = 2\pi f \qquad \text{and} \qquad d\omega = 2\pi\, df$$

Therefore, the direct and the inverse transforms are given by

$$G(2\pi f) = \int_{-\infty}^{\infty} g(t)e^{-j2\pi f t}dt \qquad \text{and} \qquad g(t) = \int_{-\infty}^{\infty} G(2\pi f)e^{j2\pi f t}df$$

This leaves only one significant difference, that of sign change in the exponential index. Otherwise the two operations are symmetrical.

$$g(t)e^{j\omega_0 t} \Longleftrightarrow G(\omega - \omega_0)$$

Observe the role reversal of time and frequency in these two equations (with the minor difference of the sign change in the exponential index). The value of this principle lies in the fact that *whenever we derive any result, we can be sure that it has a dual.* This can give valuable insights about many unsuspected properties or results in signal processing.

The properties of the Fourier transform are useful not only in deriving the direct and the inverse transforms of many functions, but also in obtaining several valuable results in signal processing. The reader should not fail to observe the ever-present duality in this discussion. We begin with the symmetry property, which is one of the consequences of the duality principle discussed.

3.3.2 Symmetry Property

This property states that if

$$g(t) \Longleftrightarrow G(\omega)$$

then

$$G(t) \Longleftrightarrow 2\pi g(-\omega) \tag{3.24}$$

Proof: From Eq. (3.8b),

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(x)e^{jxt}\,dx$$

Hence,

$$2\pi g(-t) = \int_{-\infty}^{\infty} G(x)e^{-jxt}\,dx$$

Changing t to ω yields Eq. (3.24).

EXAMPLE 3.8 In this example we shall apply the symmetry property [Eq. (3.24)] to the pair in Fig. 3.16a.

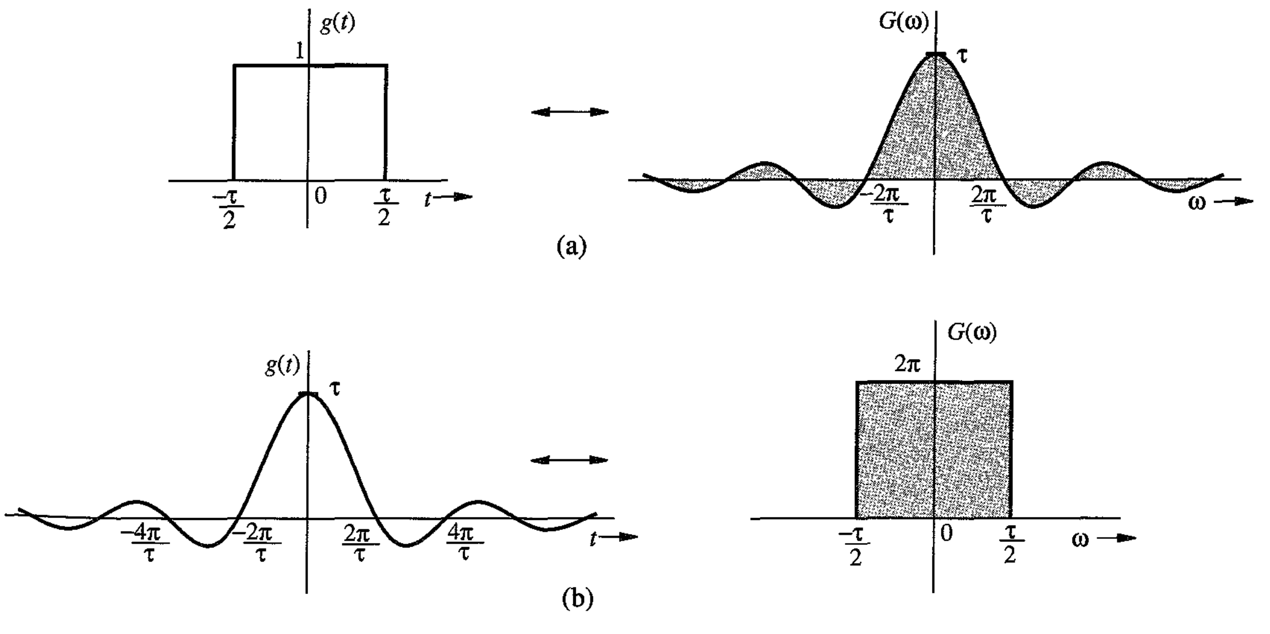


Figure 3.16 Duality property of the Fourier transform.

From Eq. (3.17) we have

$$\underbrace{\text{rect}\left(\frac{t}{\tau}\right)}_{g(t)} \Longleftrightarrow \underbrace{\tau\ \text{sinc}\left(\frac{\omega\tau}{2}\right)}_{G(\omega)} \tag{3.25}$$

Also $G(t)$ is the same as $G(\omega)$ with ω replaced by t, and $g(-\omega)$ is the same as $g(t)$ with t replaced by $-\omega$. Therefore, the symmetry property (3.24) yields

$$\underbrace{\tau\ \text{sinc}\left(\frac{\tau t}{2}\right)}_{G(t)} \Longleftrightarrow \underbrace{2\pi\ \text{rect}\left(\frac{-\omega}{\tau}\right)}_{2\pi\ g(-\omega)} = 2\pi\ \text{rect}\left(\frac{\omega}{\tau}\right) \tag{3.26}$$

In Eq. (3.26) we used the fact that rect $(-x) =$ rect (x) because rect is an even function. Figure 3.16b shows this pair graphically. Observe the interchange of the roles of t and ω (with the minor adjustment of the factor 2π). This result appears as pair 18 in Table 3.1 (with $\tau/2 = W$).

As an interesting exercise, by applying the symmetry property, the reader should generate a dual of every pair in Table 3.1.

3.3.3 Scaling Property

If

$$g(t) \Longleftrightarrow G(\omega)$$

then, for any real constant a,

$$g(at) \Longleftrightarrow \frac{1}{|a|} G\left(\frac{\omega}{a}\right) \tag{3.27}$$

Proof: For a positive real constant a,

$$\mathcal{F}[g(at)] = \int_{-\infty}^{\infty} g(at)e^{-j\omega t}\,dt = \frac{1}{a}\int_{-\infty}^{\infty} g(x)e^{(-j\omega/a)x}\,dx = \frac{1}{a}G\left(\frac{\omega}{a}\right)$$

Similarly, it can be shown that if $a < 0$,

$$g(at) \Longleftrightarrow \frac{-1}{a} G\left(\frac{\omega}{a}\right)$$

Hence follows Eq. (3.27).

Significance of the Scaling Property

The function $g(at)$ represents the function $g(t)$ compressed in time by a factor a (see Sec. 2.3.2). Similarly, a function $G(\omega/a)$ represents the function $G(\omega)$ expanded in frequency by the same factor a. *The scaling property states that time compression of a signal results in its spectral expansion, and time expansion of the signal results in its spectral compression.* Intuitively compression in time by a factor a means that the signal is varying rapidly by the same factor. To synthesize such a signal, the frequencies of its sinusoidal components must be increased by the factor a, implying that its frequency spectrum is expanded by the factor a. Similarly, a signal expanded in time varies more slowly; hence, the frequencies of its components are lowered, implying that its frequency spectrum is compressed. For instance, the signal $\cos 2\omega_0 t$ is the same as the signal $\cos \omega_0 t$ time-compressed by a factor of 2. Clearly, the spectrum of the former (impulse at $\pm 2\omega_0$) is an expanded version of the spectrum of the latter (impulse at $\pm\omega_0$). The effect of this scaling is demonstrated in Fig. 3.17.

Reciprocity of Signal Duration and Its Bandwidth

The scaling property implies that if $g(t)$ is wider, its spectrum is narrower, and vice versa. Doubling the signal duration halves its bandwidth, and vice versa. This suggests that the bandwidth of a signal is inversely proportional to the signal duration or width (in seconds). We have already verified this fact for the gate pulse, where we found that the bandwidth of a gate pulse of width τ seconds is $1/\tau$ Hz. More discussion of this interesting topic can be found in the literature.[2]

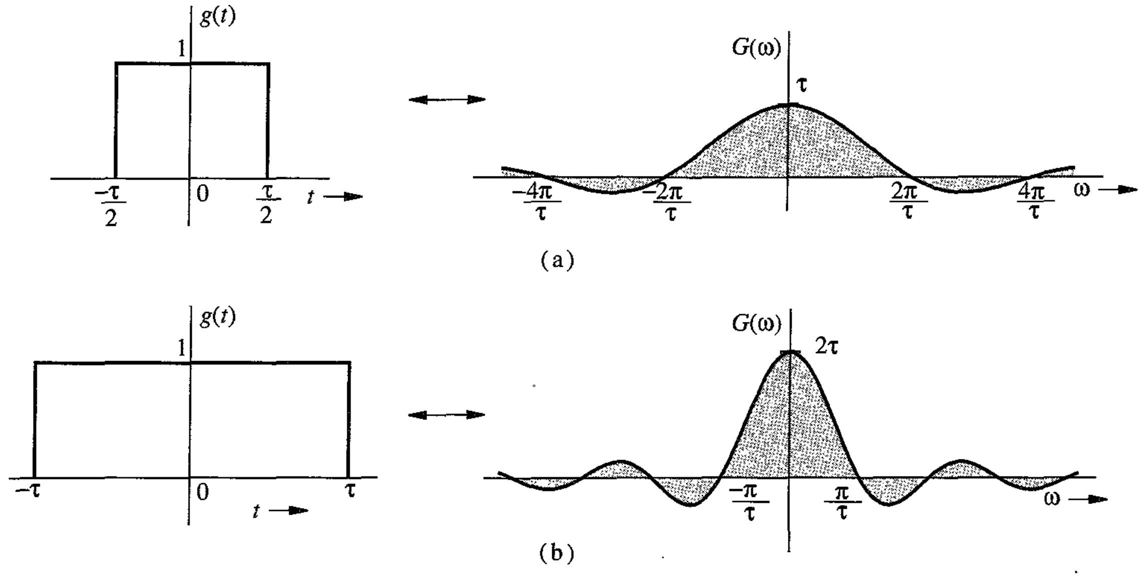
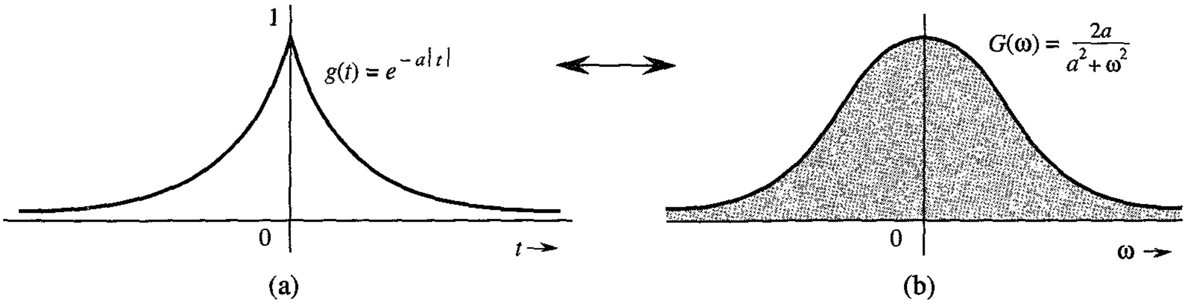


Figure 3.17 Scaling property of the Fourier transform.

EXAMPLE 3.9 Show that

$$g(-t) \Longleftrightarrow G(-\omega) \qquad (3.28)$$

Using this result and the fact that $e^{-at}u(t) \Longleftrightarrow 1/a + j\omega$, find the Fourier transforms of $e^{at}u(-t)$ and $e^{-a|t|}$.



(a) (b)

Figure 3.18 $e^{-a|t|}$ and its Fourier spectrum.

Equation (3.28) follows from Eq. (3.27) by letting $a = -1$. Application of Eq. (3.28) to pair 1 of Table 3.1 yields

$$e^{at}u(-t) \Longleftrightarrow \frac{1}{a - j\omega}$$

Also

$$e^{-a|t|} = e^{-at}u(t) + e^{at}u(-t)$$

Therefore,

$$e^{-a|t|} \Longleftrightarrow \frac{1}{a + j\omega} + \frac{1}{a - j\omega} = \frac{2a}{a^2 + \omega^2} \tag{3.29}$$

The signal $e^{-a|t|}$ and its spectrum are shown in Fig. 3.18.

3.3.4 Time-Shifting Property

If

$$g(t) \Longleftrightarrow G(\omega)$$

then

$$g(t - t_0) \Longleftrightarrow G(\omega)e^{-j\omega t_0} \tag{3.30a}$$

Proof: By definition,

$$\mathcal{F}[g(t - t_0)] = \int_{-\infty}^{\infty} g(t - t_0)e^{-j\omega t}\, dt$$

Letting $t - t_0 = x$, we have

$$\begin{aligned} \mathcal{F}[g(t - t_0)] &= \int_{-\infty}^{\infty} g(x)e^{-j\omega(x + t_0)}\, dx \\ &= e^{-j\omega t_0} \int_{-\infty}^{\infty} g(x)e^{-j\omega x}\, dx = G(\omega)e^{-j\omega t_0} \end{aligned} \tag{3.30b}$$

This result shows that *delaying a signal by t_0 seconds does not change its amplitude spectrum. The phase spectrum, however, is changed by $-\omega t_0$.*

Physical Explanation of the Linear Phase

Time delay in a signal causes a linear phase shift in its spectrum. This result can also be derived by heuristic reasoning. Imagine $g(t)$ being synthesized by its Fourier components, which are sinusoids of certain amplitudes and phases. The delayed signal $g(t - t_0)$ can be synthesized by the same sinusoidal components, each delayed by t_0 seconds. The amplitudes of the components remain unchanged. Therefore, the amplitude spectrum of $g(t - t_0)$ is identical to that of $g(t)$. The time delay of t_0 in each sinusoid, however, does change the phase of each component. Now, a sinusoid $\cos \omega t$ delayed by t_0 is given by

$$\cos \omega(t - t_0) = \cos (\omega t - \omega t_0)$$

Therefore, a time delay t_0 in a sinusoid of frequency ω manifests as a phase delay of ωt_0. This is a linear function of ω, meaning that higher frequency components must undergo proportionately higher phase shifts to achieve the same time delay. This effect is shown in Fig. 3.19 with two sinusoids, the frequency of the lower sinusoid being twice that of the upper. The same time delay t_0 amounts to a phase shift of $\pi/2$ in the upper sinusoid and a phase shift of π in the lower sinusoid. This verifies the fact that *to achieve the same time delay, higher frequency sinusoids must undergo proportionately higher phase shifts.*

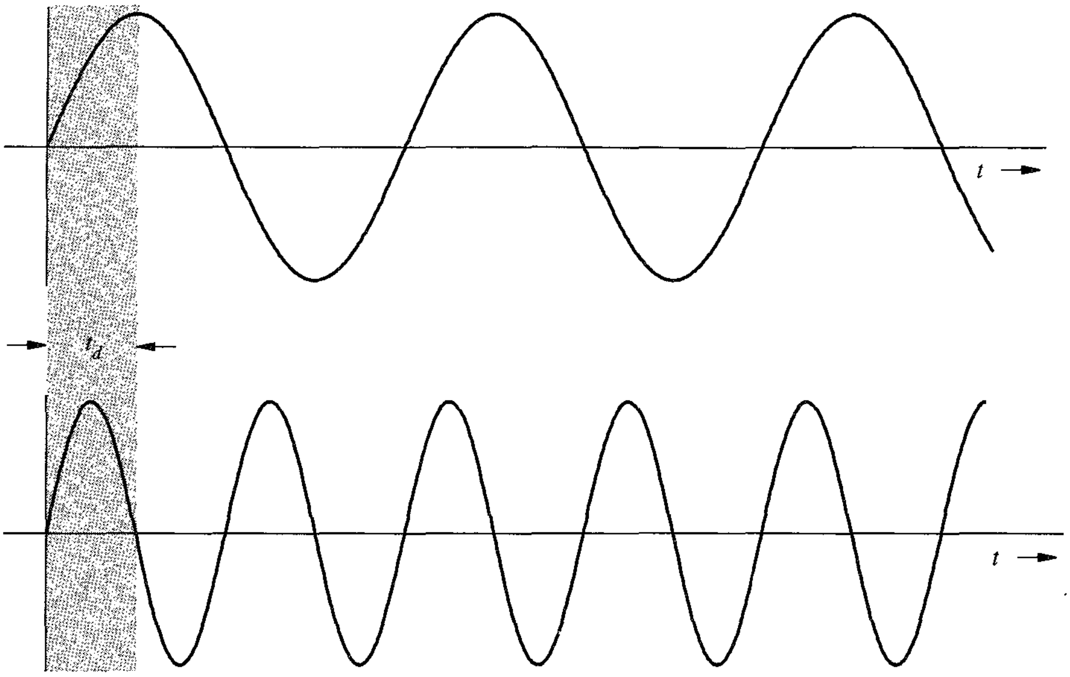


Figure 3.19 Physical explanation of the time-shifting property.

EXAMPLE 3.10 Find the Fourier transform of $e^{-a|t-t_0|}$.

This function, shown in Fig. 3.20a, is a time-shifted version of $e^{-a|t|}$ (shown in Fig. 3.18a). From Eqs. (3.29) and (3.30) we have

$$e^{-a|t-t_0|} \Longleftrightarrow \frac{2a}{a^2+\omega^2}e^{-j\omega t_0} \tag{3.31}$$

The spectrum of $e^{-a|t-t_0|}$ (Fig. 3.20b) is the same as that of $e^{-a|t|}$ (Fig. 3.18b), except for an added phase shift of $-\omega t_0$.

Observe that the time delay t_0 causes a linear phase spectrum $-\omega t_0$. This example clearly demonstrates the effect of time shift.

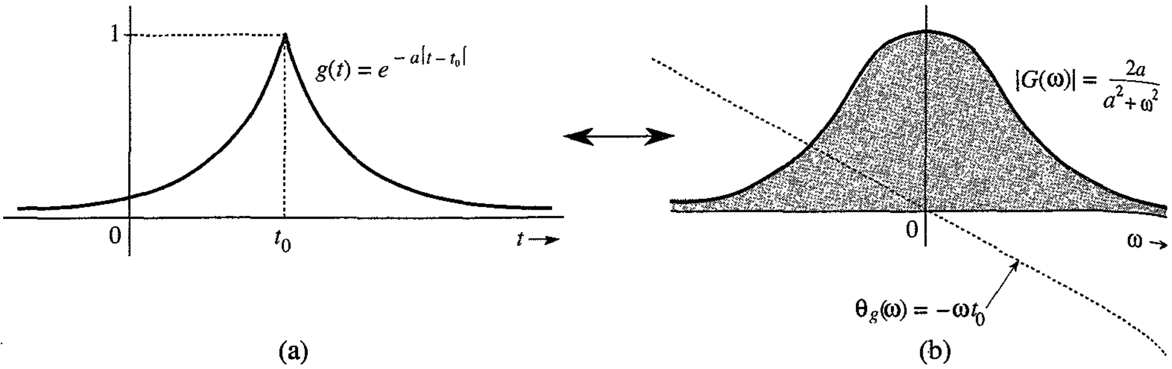


Figure 3.20 Effect of time shifting on the Fourier spectrum of a signal.

EXAMPLE 3.11 Show that

$$g(t-T)+g(t+T) \Longleftrightarrow 2G(\omega)\ \cos T\omega \tag{3.32}$$

This follows directly from Eqs. (3.30).

3.3.5 Frequency-Shifting Property

If

$$g(t) \Longleftrightarrow G(\omega)$$

then

$$g(t)e^{j\omega_0 t} \Longleftrightarrow G(\omega-\omega_0) \tag{3.33}$$

Proof: By definition,

$$\mathcal{F}[g(t)e^{j\omega_0 t}] = \int_{-\infty}^{\infty} g(t)e^{j\omega_0 t}e^{-j\omega t}dt = \int_{-\infty}^{\infty} g(t)e^{-j(\omega-\omega_0)t}dt = G(\omega-\omega_0)$$

This property states that multiplication of a signal by a factor $e^{j\omega_0 t}$ shifts the spectrum of that signal by $\omega = \omega_0$. Note the duality between the time-shifting and the frequency-shifting properties.

Changing ω_0 to $-\omega_0$ in Eq. (3.33) yields

$$g(t)e^{-j\omega_0 t} \Longleftrightarrow G(\omega + \omega_0) \tag{3.34}$$

Because $e^{j\omega_0 t}$ is not a real function that can be generated, frequency shifting in practice is achieved by multiplying $g(t)$ by a sinusoid. This can be seen from the fact that

$$g(t) \cos \omega_0 t = \frac{1}{2}\left[g(t)e^{j\omega_0 t} + g(t)e^{-j\omega_0 t}\right]$$

From Eqs. (3.33) and (3.34), it follows that

$$g(t) \cos \omega_0 t \Longleftrightarrow \frac{1}{2}\left[G(\omega - \omega_0) + G(\omega + \omega_0)\right] \tag{3.35}$$

This shows that the multiplication of a signal $g(t)$ by a sinusoid of frequency ω_0 shifts the spectrum $G(\omega)$ by $\pm\omega_0$. Multiplication of a sinusoid cos $\omega_0 t$ by $g(t)$ amounts to modulating the sinusoid amplitude. This type of modulation is known as **amplitude modulation**. The sinusoid cos $\omega_0 t$ is called the **carrier**, the signal $g(t)$ is the **modulating signal**, and the signal $g(t)$ cos $\omega_0 t$ is the **modulated signal**. Modulation and demodulation will be discussed in detail in Chapter 4.

To sketch a signal $g(t)$ cos $\omega_0 t$, we observe that

$$g(t) \cos \omega_0 t = \begin{cases} g(t) & \text{when } \cos \omega_0 t = 1 \\ -g(t) & \text{when } \cos \omega_0 t = -1 \end{cases}$$

Therefore, $g(t)$ cos $\omega_0 t$ touches $g(t)$ when the sinusoid cos $\omega_0 t$ is at its positive peaks and touches $-g(t)$ when cos $\omega_0 t$ is at its negative peaks. This means that $g(t)$ and $-g(t)$ act as envelopes for the signal $g(t)$ cos $\omega_0 t$ (see Fig. 3.21c). The signal $-g(t)$ is a mirror image of $g(t)$ about the horizontal axis. Figure 3.21 shows the signals $g(t)$, $g(t)$ cos $\omega_0 t$ and their spectra.

Shifting the Phase Spectrum of a Modulated Signal

We can shift the phase of each spectral component of a modulated signal by a constant amount θ_0 merely by using a carrier cos $(\omega_0 t + \theta_0)$ instead of cos $\omega_0 t$. If a signal $g(t)$ is multiplied by cos $(\omega_0 t + \theta_0)$, then using an argument similar to that used to derive Eq. (3.35), we can show that

$$g(t) \cos (\omega_0 t + \theta_0) \Longleftrightarrow \frac{1}{2}\left[G(\omega - \omega_0)\, e^{j\theta_0} + G(\omega + \omega_0)\, e^{-j\theta_0}\right] \tag{3.36}$$

For a special case when $\theta_0 = -\pi/2$, Eq. (3.36) becomes

$$g(t) \sin \omega_0 t \Longleftrightarrow \frac{1}{2}\left[G(\omega - \omega_0)\, e^{-j\pi/2} + G(\omega + \omega_0)\, e^{j\pi/2}\right] \tag{3.37}$$

Observe that sin $\omega_0 t$ is cos $\omega_0 t$ with a phase delay of $\pi/2$. Thus, shifting the carrier phase by $\pi/2$ shifts the phase of every spectral component by $\pi/2$. Figure 3.21e and f shows the signal $g(t)$ sin $\omega_0 t$ and its spectrum.

EXAMPLE 3.12 Find and sketch the Fourier transform of the modulated signal $g(t)$ cos $\omega_0 t$ in which $g(t)$ is a gate pulse rect (t/T), as shown in Fig. 3.22a.

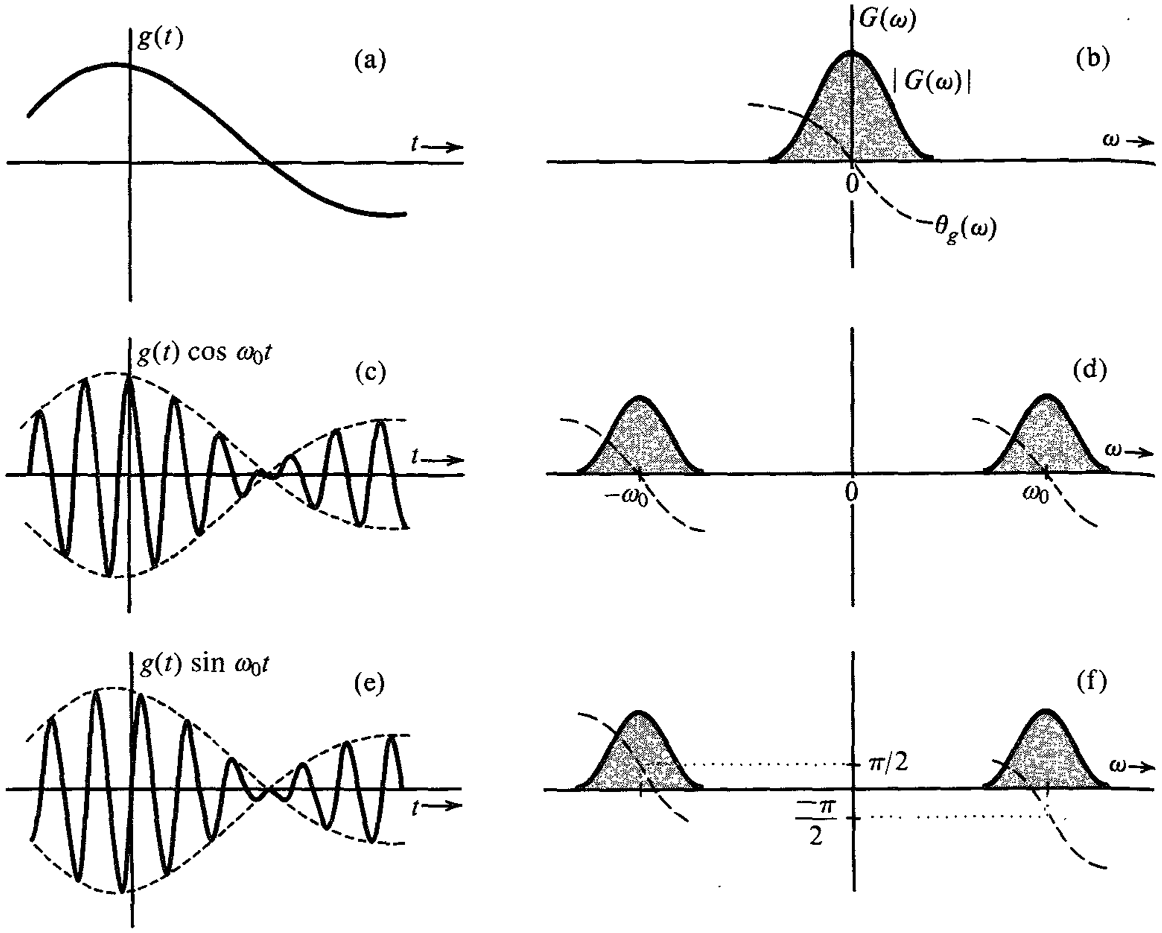

$g(t)$
(a)
$t \rightarrow$
$G(\omega)$
$|G(\omega)|$
(b)
$\omega \rightarrow$
0
$\theta_g(\omega)$
$g(t)\cos\omega_0 t$
(c)
$t \rightarrow$
(d)
$\omega \rightarrow$
$-\omega_0$
0
ω_0
$g(t)\sin\omega_0 t$
(e)
$t \rightarrow$
(f)
$\pi/2$
$\omega \rightarrow$
$\frac{-\pi}{2}$

Figure 3.21 Amplitude modulation of a signal causes spectral shifting.

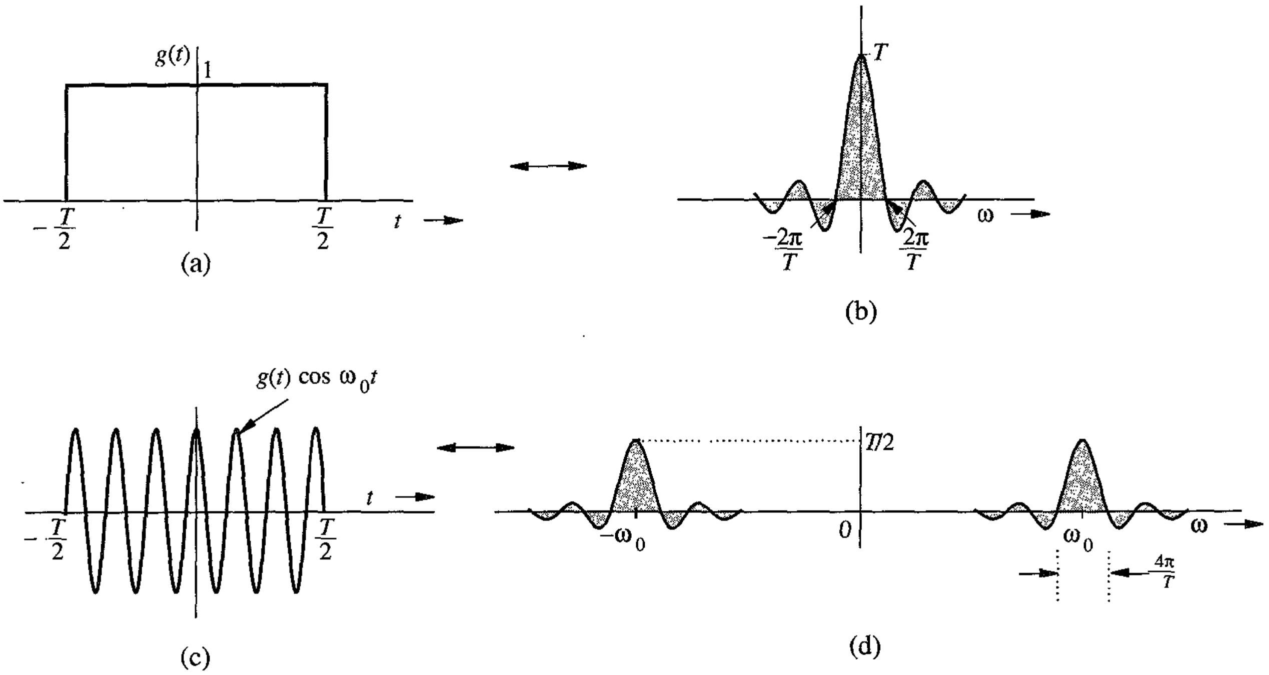

$g(t)$
1
$-\frac{T}{2}$
$\frac{T}{2}$
$t \rightarrow$
(a)
T
$\frac{-2\pi}{T}$
$\frac{2\pi}{T}$
$\omega \rightarrow$
(b)
$g(t)\cos\omega_0 t$
$t \rightarrow$
$-\frac{T}{2}$
$\frac{T}{2}$
(c)
$T/2$
$-\omega_0$
0
ω_0
$\omega \rightarrow$
$\frac{4\pi}{T}$
(d)

Figure 3.22 Example of spectral shifting by amplitude modulation.

The pulse $g(t)$ is the same rectangular pulse shown in Fig. 3.10a (with $\tau = T$). From pair 17 of Table 3.1, we find $G(\omega)$, the Fourier transform of $g(t)$, as

$$\text{rect}\left(\frac{t}{T}\right) \Longleftrightarrow T \,\text{sinc}\left(\frac{\omega T}{2}\right)$$

This spectrum $G(\omega)$ is shown in Fig. 3.22b. The signal $g(t)\cos\omega_0 t$ is shown in Fig. 3.22c. From Eq. (3.35) it follows that

$$g(t)\cos\omega_0 t \Longleftrightarrow \frac{1}{2}[G(\omega+\omega_0) + G(\omega-\omega_0)]$$

This spectrum of $g(t)\cos\omega_0 t$ is obtained by shifting $G(\omega)$ in Fig. 3.22b to the left by ω_0 and also to the right by ω_0 and then multiplying it by half, as shown in Fig. 3.22d.

Application of Modulation

Modulation is used to shift signal spectra. Some of the situations where spectrum shifting is necessary are given next.

1. If several signals, each occupying the same frequency band, are transmitted simultaneously over the same transmission medium, they will all interfere; it will be impossible to separate or retrieve them at a receiver. For example, if all radio stations decide to broadcast audio signals simultaneously, the receiver will not be able to separate them. This problem is solved by using modulation, whereby each radio station is assigned a distinct carrier frequency. Each station transmits a modulated signal, thus shifting the signal spectrum to its allocated band, which is not occupied by any other station. A radio receiver can pick up any station by tuning to the band of the desired station. The receiver must now demodulate the received signal (undo the effect of modulation). Demodulation therefore consists of another spectral shift required to restore the signal to its original band. Note that both modulation and demodulation implement spectral shifting. Consequently, demodulation operation is similar to modulation (see Prob. 3.3-10). This method of transmitting several signals simultaneously over a channel by sharing its frequency band is known as **frequency-division multiplexing (FDM)**.
2. For effective radiation of power over a radio link, the antenna size must be on the order of the wavelength of the signal to be radiated. Audio signal frequencies are so low (wavelengths are so large) that impracticably large antennas will be required for radiation. Here, shifting the spectrum to a higher frequency (a smaller wavelength) by modulation solves the problem.

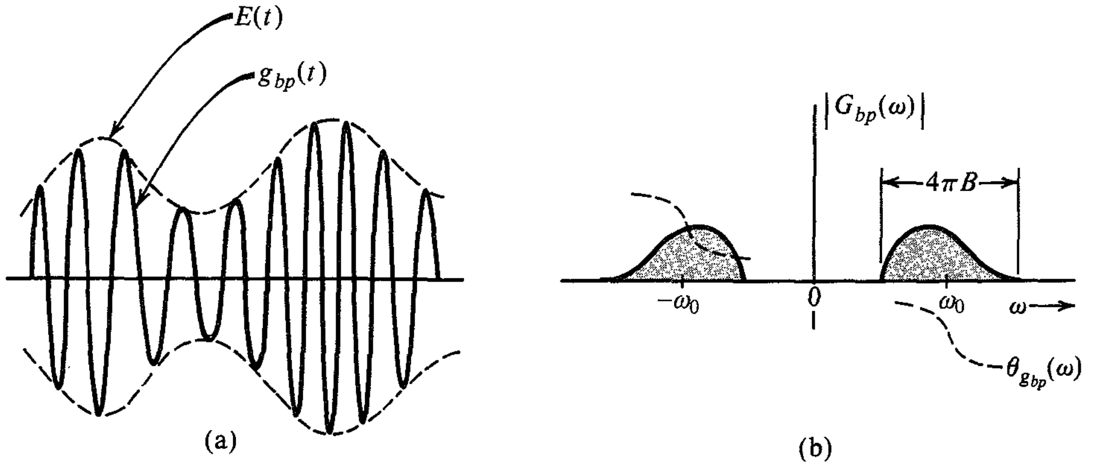
Bandpass Signals

Figure 3.21d and f shows that if $g_c(t)$ and $g_s(t)$ are low-pass signals, each with a bandwidth B Hz or $2\pi B$ rad/s, then the signals $g_c(t)\cos\omega_0 t$ and $g_s(t)\sin\omega_0 t$ are both bandpass signals occupying the same band, and each having a bandwidth of $4\pi B$ rad/s. Hence, a linear combination of both these signals will also be a bandpass signal occupying the same band as that of the either signal, and with the same bandwidth ($4\pi B$ rad/s). Hence, a general bandpass signal $g_{bp}(t)$ can be expressed as*

$$g_{bp}(t) = g_c(t)\cos\omega_0 t + g_s(t)\sin\omega_0 t \tag{3.38}$$

---

* See Sec. 11.5 for a rigorous proof of this statement.



(a) (b)

Figure 3.23 Bandpass signal and its spectrum.

The spectrum of $g_{bp}(t)$ is centered at $\pm\omega_0$ and has a bandwidth $4\pi B$, as shown in Fig. 3.23. Although the magnitude spectra of both $g_c(t)\cos\omega_0 t$ and $g_s(t)\sin\omega_0 t$ are symmetrical about $\pm\omega_0$, the magnitude spectrum of their sum, $g_{bp}(t)$, is not necessarily symmetrical about $\pm\omega_0$. This is due to the fact that the amplitudes of the two signals do not add directly because of their phases for the reason that

$$a_1 e^{j\varphi_1} + a_2 e^{j\varphi_2} \neq (a_1 + a_2)e^{j(\varphi_1+\varphi_2)}$$

A typical bandpass signal $g_{bp}(t)$ and its spectra are shown in Fig. 3.23. Using a well-known trigonometric identity, Eq. (3.38) can be expressed as

$$g_{bp}(t) = E(t)\cos[\omega_0 t + \psi(t)] \tag{3.39}$$

where

$$E(t) = +\sqrt{g_c^2(t) + g_s^2(t)} \tag{3.40a}$$

$$\psi(t) = -\tan^{-1}\left[\frac{g_s(t)}{g_c(t)}\right] \tag{3.40b}$$

Because $g_c(t)$ and $g_s(t)$ are low-pass signals, $E(t)$ and $\psi(t)$ are also low-pass signals. Because $E(t)$ is nonnegative [Eq. (3.40a)], it follows from Eq. (3.39) that $E(t)$ is a slowly varying envelope and $\psi(t)$ is a slowly varying phase of the bandpass signal $g_{bp}(t)$, as shown in Fig. 3.23. Thus, the bandpass signal $g_{bp}(t)$ will appear as a sinusoid of slowly varying amplitude. Because of the time-varying phase $\psi(t)$ the frequency of the sinusoid also varies slowly* with time about the center frequency ω_0.

EXAMPLE 3.13 Find the Fourier transform of a general periodic signal $g(t)$ of period T_0, and hence, determine the Fourier transform of the periodic impulse train $\delta_{T_0}(t)$ shown in Fig. 3.24a.

---

* It is necessary that $2\pi B << \omega_0$ for a well-defined envelope. Otherwise the variations of $E(t)$ are of the same order as the carrier, and it will be difficult to separate the envelope from the carrier.

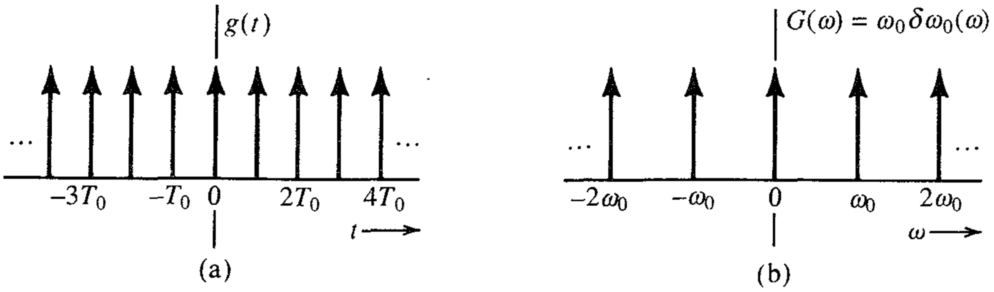


Figure 3.24 Impulse train and its spectrum.

A periodic signal $g(t)$ can be expressed as an exponential Fourier series as

$$g(t) = \sum_{n=-\infty}^{\infty} D_n\, e^{jn\omega_0 t} \qquad \omega_0 = \frac{2\pi}{T_0}$$

Therefore,

$$g(t) \Longleftrightarrow \sum_{n=-\infty}^{\infty} \mathcal{F}[D_n\, e^{jn\omega_0 t}]$$

Now from Eq. (3.20a), it follows that

$$g(t) \Longleftrightarrow 2\pi \sum_{n=-\infty}^{\infty} D_n\, \delta(\omega - n\omega_0) \tag{3.41}$$

Equation (2.89) shows that the impulse train $\delta_{T_0}(t)$ can be expressed as an exponential Fourier series as

$$\delta_{T_0}(t) = \frac{1}{T_0} \sum_{n=-\infty}^{\infty} e^{jn\omega_0 t} \qquad \omega_0 = \frac{2\pi}{T_0}$$

Here $D_n = 1/T_0$. Therefore, from Eq. (3.41),

$$\begin{aligned} \delta_{T_0}(t) &\Longleftrightarrow \frac{2\pi}{T_0} \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \\ &= \omega_0 \delta_{\omega_0}(\omega) \qquad \omega_0 = \frac{2\pi}{T_0} \end{aligned} \tag{3.42}$$

Thus, the spectrum of the impulse train also happens to be an impulse train (in the frequency domain), as shown in Fig. 3.24b.

3.3.6 Convolution

The convolution of two functions $g(t)$ and $w(t)$, denoted by $g(t) * w(t)$, is defined by the integral

$$g(t) * w(t) = \int_{-\infty}^{\infty} g(\tau) w(t - \tau)\, d\tau$$

The time convolution property and its dual, the frequency convolution property, state that if

$$g_1(t) \Longleftrightarrow G_1(\omega) \qquad \text{and} \qquad g_2(t) \Longleftrightarrow G_2(\omega)$$

then (**time convolution**)

$$g_1(t) * g_2(t) \Longleftrightarrow G_1(\omega)G_2(\omega) \tag{3.43}$$

and (**frequency convolution**)

$$g_1(t)g_2(t) \Longleftrightarrow \frac{1}{2\pi}G_1(\omega) * G_2(\omega) \tag{3.44}$$

Proof: By definition,

$$\begin{aligned} \mathcal{F}|g_1(t) * g_2(t)| &= \int_{-\infty}^{\infty} e^{-j\omega t}\left[\int_{-\infty}^{\infty} g_1(\tau)g_2(t-\tau)d\tau\right] dt \\ &= \int_{-\infty}^{\infty} g_1(\tau)\left[\int_{-\infty}^{\infty} e^{-j\omega t}g_2(t-\tau)dt\right] d\tau \end{aligned}$$

The inner integral is the Fourier transform of $g_2(t-\tau)$, given by [time-shifting property in Eq. (3.30)] $G_2(\omega)e^{-j\omega\tau}$. Hence,

$$\begin{aligned} \mathcal{F}[g_1(t) * g_2(t)] &= \int_{-\infty}^{\infty} g_1(\tau)e^{-j\omega\tau}G_2(\omega)\, d\tau \\ &= G_2(\omega)\int_{-\infty}^{\infty} g_1(\tau)e^{-j\omega\tau}\, d\tau = G_1(\omega)G_2(\omega) \end{aligned}$$

The frequency convolution property (3.44) can be proved in exactly the same way by reversing the roles of $g(t)$ and $G(\omega)$.

Bandwidth of the Product of Two Signals

If $g_1(t)$ and $g_2(t)$ have bandwidths B_1 and B_2 Hz, respectively, the bandwidth of $g_1(t)g_2(t)$ is $B_1 + B_2$ Hz. This result follows from the application of the width property of convolution[3] to Eq. (3.44). This property states that the width of $x * y$ is the sum of the widths of x and y. Consequently, if the bandwidth of $g(t)$ is B Hz, then the bandwidth of $g^2(t)$ is $2B$ Hz, and the bandwidth of $g^n(t)$ is nB Hz.*

EXAMPLE 3.14 Using the time convolution property, show that if

$$g(t) \Longleftrightarrow G(\omega)$$

then

$$\int_{-\infty}^{t} g(\tau)d\tau \Longleftrightarrow \frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega) \tag{3.45}$$

---

* The width property of convolution does not hold in some pathological cases. This happens when the convolution of two functions is zero over a range even when both functions are nonzero, e.g., $\sin\omega_0 t\; u(t) * u(t)$. Technically the property holds even in this case if in calculating the width of the convolved function we take into account that range where the convolution is zero.

Because

$$u(t-\tau) = \begin{cases} 1 & \tau \le t \\ 0 & \tau > t \end{cases}$$

it follows that

$$g(t) * u(t) = \int_{-\infty}^{\infty} g(\tau)u(t-\tau)d\tau = \int_{-\infty}^{t} g(\tau)\, d\tau$$

Now from the time convolution property [Eq. (3.43)], it follows that

$$\begin{aligned} g(t) * u(t) &\Longleftrightarrow G(\omega)U(\omega) \\ &= G(\omega)\left[\frac{1}{j\omega} + \pi\delta(\omega)\right] \\ &= \frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega) \end{aligned}$$

In deriving the last result we used pair 11 of Table 3.1 and Eq. (2.18a).

3.3.7 Time Differentiation and Time Integration

If

$$g(t) \Longleftrightarrow G(\omega)$$

then (**time differentiation**)*

$$\frac{dg}{dt} \Longleftrightarrow j\omega G(\omega) \tag{3.46}$$

and (**time integration**)

$$\int_{-\infty}^{t} g(\tau)d\tau \Longleftrightarrow \frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega) \tag{3.47}$$

Proof: Differentiation of both sides of Eq. (3.8b) yields

$$\frac{dg}{dt} = \frac{1}{2\pi}\int_{-\infty}^{\infty} j\omega G(\omega)e^{j\omega t}\, d\omega$$

This shows that

$$\frac{dg}{dt} \Longleftrightarrow j\omega G(\omega)$$

Repeated application of this property yields

$$\frac{d^n g}{dt^n} \Longleftrightarrow (j\omega)^n G(\omega) \tag{3.48}$$

The time integration property [Eq. (3.47)] already has been proved in Example 3.14.

---

* Valid only if the transform of dg/dt exists.

EXAMPLE 3.15 Using the time differentiation property, find the Fourier transform of the triangle pulse $\Delta(t/\tau)$ shown in Fig. 3.25a.

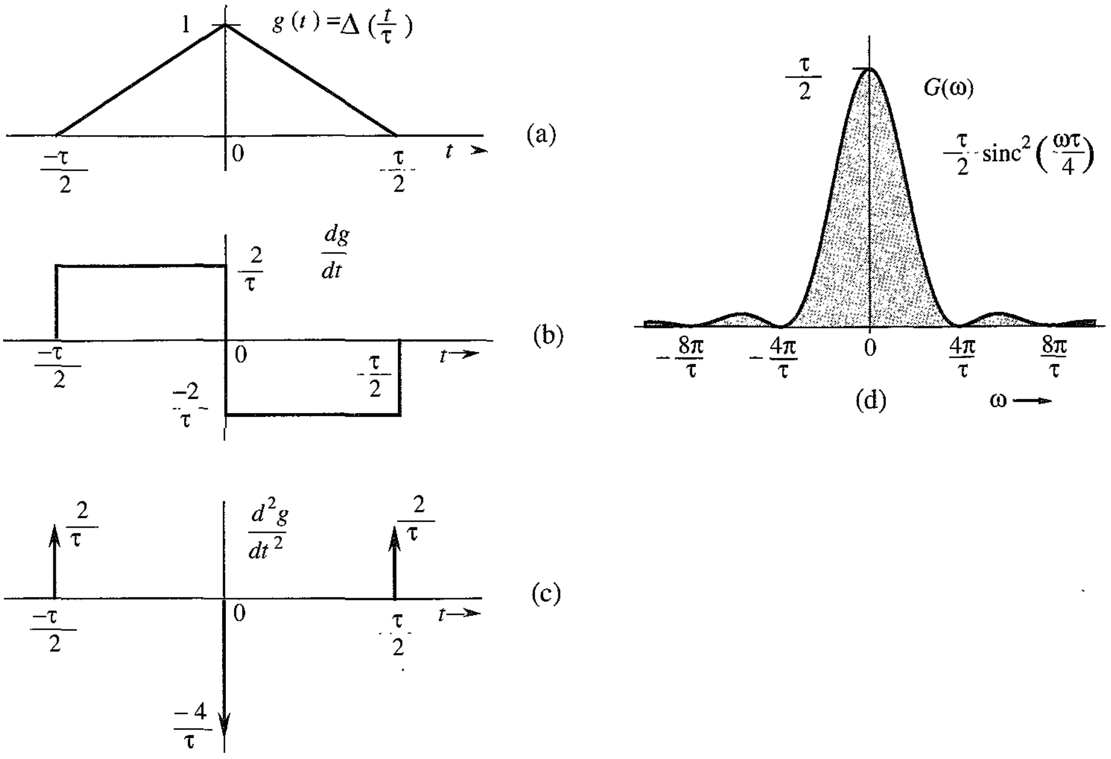


Figure 3.25 Finding the Fourier transform of a piecewise-linear signal using the time differentiation property.

To find the Fourier transform of this pulse we differentiate it successively, as shown in Fig. 3.25b and c. The second derivative consists of a sequence of impulses, as shown in Fig. 3.25c. Recall that the derivative of a signal at a jump discontinuity is an impulse of strength equal to the amount of jump. The function dg/dt has a positive jump of $2/\tau$ at $t = \pm\tau/2$, and a negative jump of $4/\tau$ at $t = 0$. Therefore,

$$\frac{d^2g}{dt^2} = \frac{2}{\tau}\left[\delta\left(t + \frac{\tau}{2}\right) - 2\delta(t) + \delta\left(t - \frac{\tau}{2}\right)\right] \tag{3.49}$$

From the time differentiation property (3.48),

$$\frac{d^2g}{dt^2} \Longleftrightarrow (j\omega)^2 G(\omega) = -\omega^2 G(\omega) \tag{3.50a}$$

Also, from the time-shifting property (3.30),

$$\delta(t - t_0) \Longleftrightarrow e^{-j\omega t_0} \tag{3.50b}$$

Taking the Fourier transform of Eq. (3.49) and using the results in Eqs. (3.50), we obtain

$$-\omega^2 G(\omega) = \frac{2}{\tau}\left(e^{j\frac{\omega\tau}{2}} - 2 + e^{-j\frac{\omega\tau}{2}}\right) = \frac{4}{\tau}\left(\cos\frac{\omega\tau}{2} - 1\right) = -\frac{8}{\tau}\sin^2\left(\frac{\omega\tau}{4}\right)$$

and

$$G(\omega) = \frac{8}{\omega^2\tau}\sin^2\left(\frac{\omega\tau}{4}\right) = \frac{\tau}{2}\left[\frac{\sin(\omega\tau/4)}{\omega\tau/4}\right]^2 = \frac{\tau}{2}\,\mathrm{sinc}^2\left(\frac{\omega\tau}{4}\right) \tag{3.51}$$

The spectrum $G(\omega)$ is shown in Fig. 3.25d. This procedure of finding the Fourier transform can be applied to any function $g(t)$ made up of straight-line segments with $g(t) \to 0$ as $|t| \to \infty$. The second derivative of such a signal yields a sequence of impulses whose Fourier transform can be found by inspection. This example suggests a numerical method of finding the Fourier transform of an arbitrary signal $g(t)$ by approximating the signal by straight-line segments.

3.4 SIGNAL TRANSMISSION THROUGH A LINEAR SYSTEM

For a linear, time-invariant, continuous-time system the input-output relationship is given by

$$y(t) = g(t) * h(t) \tag{3.52}$$

where $g(t)$ is the input, $y(t)$ is the output, and $h(t)$ is the unit impulse response of the linear time-invariant system. If

$$g(t) \Longleftrightarrow G(\omega), \qquad y(t) \Longleftrightarrow Y(\omega), \qquad \text{and} \quad h(t) \Longleftrightarrow H(\omega)$$

where $H(\omega)$ is the system transfer function, then application of the time convolution property to Eq. (3.52) yields

$$Y(\omega) = G(\omega)H(\omega) \tag{3.53}$$

3.4.1 Signal Distortion during Transmission

The transmission of an input signal $g(t)$ through a system changes it into the output signal $y(t)$. Equation (3.53) shows the nature of this change or modification. Here $G(\omega)$ and $Y(\omega)$ are the

Table 3.2
Fourier Transform Operations

Operation	$g(t)$	$G(\omega)$
Addition	$g_1(t) + g_2(t)$	$G_1(\omega) + G_2(\omega)$
Scalar multiplication	$kg(t)$	$kG(\omega)$
Symmetry	$G(t)$	$2\pi g(-\omega)$
Scaling	$g(at)$	$\frac{1}{\lvert a\rvert}G\left(\frac{\omega}{a}\right)$
Time shift	$g(t - t_0)$	$G(\omega)e^{-j\omega t_0}$
Frequency shift	$g(t)e^{j\omega_0 t}$	$G(\omega - \omega_0)$
Time convolution	$g_1(t) * g_2(t)$	$G_1(\omega)G_2(\omega)$
Frequency convolution	$g_1(t)g_2(t)$	$\frac{1}{2\pi}G_1(\omega) * G_2(\omega)$
Time differentiation	$\frac{d^n g}{dt^n}$	$(j\omega)^n G(\omega)$
Time integration	$\int_{-\infty}^{t} g(x)\,dx$	$\frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega)$

spectra of the input and the output, respectively. Therefore, $H(\omega)$ is the spectral response of the system. The output spectrum is given by the input spectrum multiplied by the spectral response of the system. Equation (3.53) clearly brings out the spectral shaping (or modification) of the signal by the system. Equation (3.53) can be expressed in polar form as

$$|Y(\omega)|e^{j\theta_y(\omega)} = |G(\omega)||H(\omega)|e^{j[\theta_g(\omega)+\theta_h(\omega)]}$$

Therefore,

$$|Y(\omega)| = |G(\omega)|\,|H(\omega)| \tag{3.54a}$$

$$\theta_y(\omega) = \theta_g(\omega) + \theta_h(\omega) \tag{3.54b}$$

During the transmission, the input signal amplitude spectrum $|G(\omega)|$ is changed to $|G(\omega)|\,|H(\omega)|$. Similarly, the input signal phase spectrum $\theta_g(\omega)$ is changed to $\theta_g(\omega) + \theta_h(\omega)$. An input signal spectral component of frequency ω is modified in amplitude by a factor $|H(\omega)|$ and is shifted in phase by an angle $\theta_h(\omega)$. Clearly, $|H(\omega)|$ is the amplitude response, and $\theta_h(\omega)$ is the phase response of the system. The plots of $|H(\omega)|$ and $\theta_h(\omega)$ as functions of ω show at a glance how the system modifies the amplitudes and phases of various sinusoidal inputs. This is why $H(\omega)$ is called the **frequency response** of the system. During transmission through the system, some frequency components may be boosted in amplitude, while others may be attenuated. The relative phases of the various components also change. In general, the output waveform will be different from the input waveform.

Distortionless Transmission

In several applications, such as signal amplification or message signal transmission over a communication channel, we require the output waveform to be a replica of the input waveform. In such cases, we need to minimize the distortion caused by the amplifier or the communication channel. It is therefore of practical interest to determine the characteristics of a system that allows a signal to pass without distortion (**distortionless transmission**).

Transmission is said to be distortionless if the input and the output have identical wave shapes within a multiplicative constant. A delayed output that retains the input waveform is also considered distortionless. Thus, in distortionless transmission, the input $g(t)$ and the output $y(t)$ satisfy the condition

$$y(t) = kg(t - t_d) \tag{3.55}$$

The Fourier transform of this equation yields

$$Y(\omega) = kG(\omega)e^{-j\omega t_d}$$

But

$$Y(\omega) = G(\omega)H(\omega)$$

Therefore,

$$H(\omega) = k\,e^{-j\omega t_d}$$

This is the transfer function required for distortionless transmission. From this equation it follows that

$$|H(\omega)| = k \tag{3.56a}$$

$$\theta_h(\omega) = -\omega t_d \tag{3.56b}$$

This shows that for distortionless transmission, the amplitude response $|H(\omega)|$ must be a constant, and the phase response $\theta_h(\omega)$ must be a linear function of ω, as shown in Fig. 3.26. The slope of $\theta_h(\omega)$ with respect to ω is $-t_d$, where t_d is the delay of the output with respect to the input.*

Intuitive Explanation of the Distortionless Transmission Conditions

It is instructive to derive the conditions for distortionless transmission heuristically. Once again, imagine $g(t)$ to be composed of various sinusoids (its spectral components), which are being passed through a distortionless system. For the distortionless case, the output signal is the input signal multiplied by k and delayed by t_d. To synthesize such a signal, we need exactly the same components as those of $g(t)$, with each component multiplied by k and delayed by t_d. This means that the system transfer function $H(\omega)$ should be such that each sinusoidal component suffers the same attenuation k and each component undergoes the same time delay of t_d seconds. The first condition requires that

$$|H(\omega)| = k$$

We have seen earlier (sec. 3.3) that to achieve the same time delay t_d for every frequency component requires a linear phase delay ωt_d (Fig. 3.19). Therefore,

$$\theta_h(\omega) = -\omega t_d$$

The time delay resulting from the signal transmission through a system is the negative of the slope of the system phase response θ_h; that is,

$$t_d(\omega) = -\frac{d\theta_h}{d\omega} \tag{3.57}$$

If the slope of θ_h is constant (that is, if θ_h is linear with ω), all the components are delayed by the same time interval t_d. But if the slope is not constant, t_d, the time delay, varies with frequency. This means that different frequency components undergo different amounts of time delay, and consequently the output waveform will not be a replica of the input waveform. A good way of judging phase linearity is to plot t_d as a function of frequency. For a distortionless system, t_d should be constant over the band of interest.†

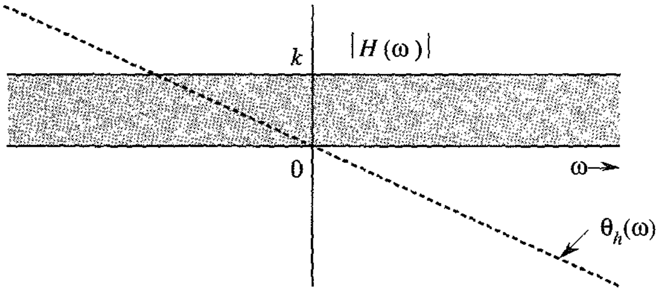


Figure 3.26 Linear time-invariant system frequency response for distortionless transmission.

---

* In addition, we require that $\theta_h(0)$ either be 0 (as shown in Fig. 3.26) or have a constant value $n\pi$ (n an integer), that is, $\theta_h(\omega) = n\pi - \omega t_d$. The addition of the excess phase of $n\pi$ may at most change the sign of the signal.

† Figure 3.26 shows that for distortionless transmission, the phase response is not only linear, but must also pass through the origin. This latter requirement can be somewhat relaxed for bandpass signals. The phase at the origin may be any constant [$\theta_h(\omega) = \theta_0 - \omega t_d$ or $\theta_h(0) = \theta_0$]. The reason for this can be found in Eq. (3.36), which shows that the addition of a constant phase θ_0 to a spectrum of a bandpass signal amounts to a phase shift of the carrier by θ_0. The modulating signal (the envelope) is not affected. The output envelope is the same as the input envelope delayed

It is often thought (erroneously) that flatness of amplitude response $|H(\omega)|$ alone can guarantee signal quality. A system may have a flat amplitude response and yet distort a signal beyond recognition if the phase response is not linear (t_d not constant).

The Nature of Distortion in Audio and Video Signals

Generally speaking, a human ear can readily perceive amplitude distortion, although it is relatively insensitive to phase distortion. For the phase distortion to become noticeable, the variation in delay (variation in the slope of θ_h) should be comparable to the signal duration (or the physically perceptible duration, in case the signal itself is long). In the case of audio signals, each spoken syllable can be considered as an individual signal. The average duration of a spoken syllable is of a magnitude on the order of 0.01 to 0.1 second. The audio systems may have nonlinear phases, yet no noticeable signal distortion results because in practical audio systems, maximum variation in the slope of θ_h is only a small fraction of a millisecond. This is the real reason behind the statement that "the human ear is relatively insensitive to phase distortion."[4] As a result, the manufacturers of audio equipment make available only $|H(\omega)|$, the amplitude response characteristic of their systems.

For video signals, on the other hand, the situation is exactly the opposite. The human eye is sensitive to phase distortion but is relatively insensitive to amplitude distortion. The amplitude distortion in television signals manifests itself as a partial destruction of the relative half-tone values of the resulting picture, which is not readily apparent to the human eye. The phase distortion (nonlinear phase), on the other hand, causes different time delays in different picture elements. This results in a smeared picture, which is readily apparent to the human eye. Phase distortion is also very important in digital communication systems because the nonlinear phase characteristic of a channel causes pulse dispersion (spreading out), which in turn causes pulses to interfere with neighboring pulses. This interference can cause an error in the pulse amplitude at the receiver: a binary **1** may read as **0**, and vice versa.

EXAMPLE 3.16 If $g(t)$ and $y(t)$ are the input and the output, respectively, of a simple RC low-pass filter (Fig. 3.27a), determine the transfer function $H(\omega)$ and sketch $|H(\omega)|$, $\theta_h(\omega)$, and $t_d(\omega)$. For distortionless transmission through this filter, what is the requirement on the bandwidth of $g(t)$ if amplitude response variation within 2% and time delay variation within 5% are tolerable? What is the transmission delay? Find the output $y(t)$.

Application of the voltage division rule to this circuit yields

$$H(\omega) = \frac{1/j\omega C}{R + (1/j\omega C)} = \frac{1}{1 + j\omega RC} = \frac{a}{a + j\omega}$$

where

$$a = \frac{1}{RC} = 10^6$$

Hence,

$$|H(\omega)| = \frac{a}{\sqrt{a^2 + \omega^2}} \simeq 1 \qquad \omega \ll a$$

$$\theta_h(\omega) = -\tan^{-1}\frac{\omega}{a} \simeq -\frac{\omega}{a} \qquad \omega \ll a$$

---

by $t_g = -d\theta_h/d\omega$, called the **group** or **envelope delay**, and the output carrier is the same as the input carrier delayed by $t_p = -\theta_h(\omega_0)/\omega_0$, called the **phase delay**, where ω_0 is the center frequency of the passband.

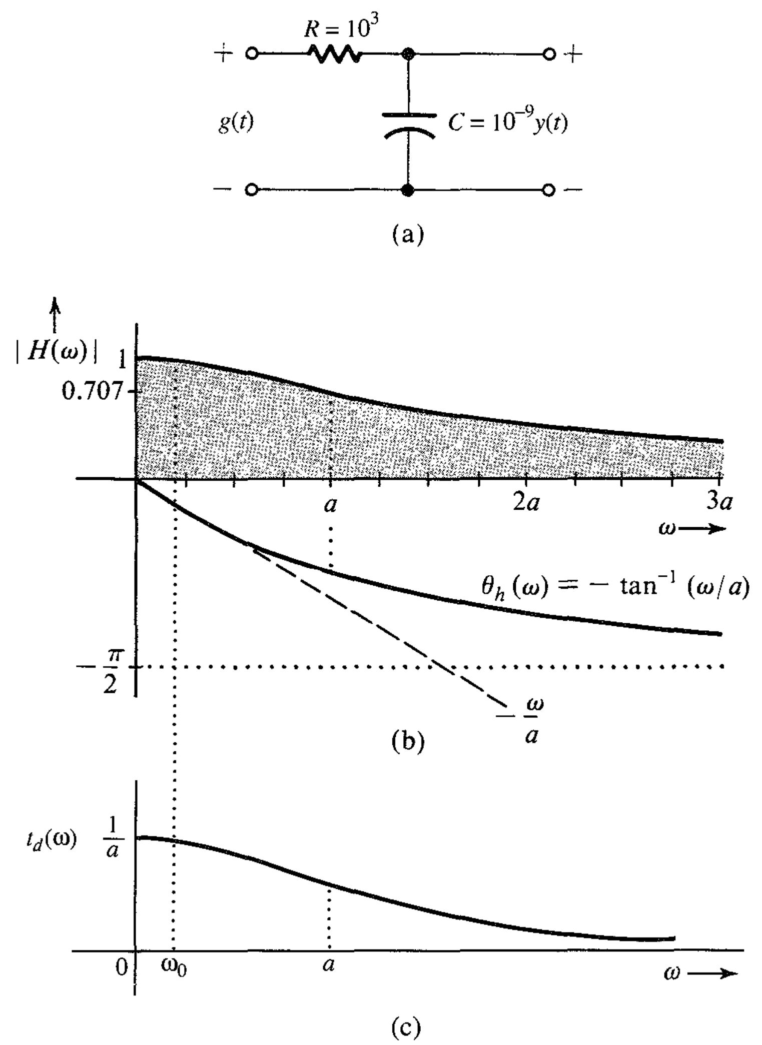


Figure 3.27 Simple RC filter, its frequency response and time delay.

Finally, the time delay is given by [Eq. (3.57)]

$$t_d(\omega) = -\frac{d\theta_h}{d\omega} = \frac{a}{\omega^2 + a^2} \simeq \frac{1}{a} = 10^{-6} \qquad \omega \ll a$$

The amplitude and phase response characteristics are given in Fig. 3.27b. The time delay t_d as a function of ω is shown in Fig. 3.27c. For $\omega \ll a$ ($a = 10^6$), the amplitude response is practically constant and the phase shift is nearly linear. The phase linearity results in a constant time delay characteristic. The filter therefore can transmit low-frequency signals with negligible distortion.

In our case, amplitude response variation within 2% and time delay variation within 5% are tolerable. Let ω_0 be the highest bandwidth of a signal that can be transmitted within these specifications. To compute ω_0 observe that the filter is a low-pass filter with gain and time delay both at maximum when $\omega = 0$ and

$$|H(0)| = 1 \qquad \text{and} \qquad t_d(0) = \frac{1}{a}$$

Therefore, $|H(\omega_0)| \geq 0.98$ and $t_d(\omega_0) \geq 0.95/a$, so that

$$|H(\omega_0)| = \frac{a}{\sqrt{\omega_0^2 + a^2}} \geq 0.98 \Longrightarrow \omega_0 \leq 0.203a = 203,000$$

$$t_d(\omega_0) = \frac{a}{\omega_0^2 + a^2} \geq \frac{0.95}{a} \Longrightarrow \omega_0 \leq 0.2294a = 229,400$$

The smaller of the two values, $\omega_0 = 203,000$ rad/s or 32.31 kHz, is the highest bandwidth that satisfies both constraints on $|H(\omega)|$ and t_d.

The time delay $t_d \approx 1/a = 10^{-6}$ over this band (see Fig. 3.27c). Also the amplitude response is almost unity (Fig. 3.27b). Therefore, the output $y(t) \approx g(t - 10^{-6})$.

3.5 IDEAL AND PRACTICAL FILTERS

Ideal filters allow distortionless transmission of a certain band of frequencies and suppress all the remaining frequencies. The ideal low-pass filter (Fig. 3.28), for example, allows all components below $\omega = W$ rad/s to pass without distortion and suppresses all components above $\omega = W$. Figure 3.29 shows ideal high-pass and bandpass filter characteristics.

The ideal low-pass filter in Fig. 3.28a has a linear phase of slope $-t_d$, which results in a time delay of t_d seconds for all its input components of frequencies below W rad/s. Therefore, if the input is a signal $g(t)$ band-limited to W rad/s, the output $y(t)$ is $g(t)$ delayed by t_d, that is,

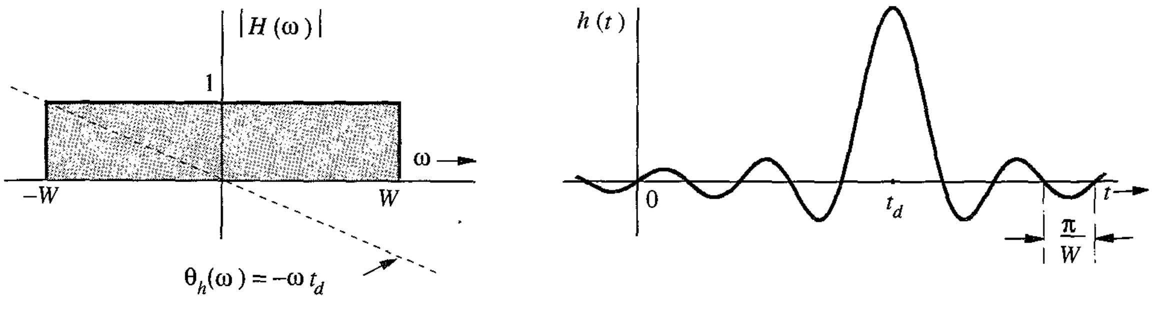
$$y(t) = g(t - t_d)$$

The signal $g(t)$ is transmitted by this system without distortion, but with time delay t_d. For this filter $|H(\omega)| = \text{rect}\,(\omega/2W)$, and $\theta_h(\omega) = -\omega t_d$, so that

$$H(\omega) = \text{rect}\left(\frac{\omega}{2W}\right) e^{-j\omega t_d} \tag{3.58a}$$

The unit impulse response $h(t)$ of this filter is found from pair 18 in Table 3.1 and the time-shifting property:

$$\begin{aligned} h(t) &= \mathcal{F}^{-1}\left[\text{rect}\left(\frac{\omega}{2W}\right) e^{-j\omega t_d}\right] \\ &= \frac{W}{\pi}\,\text{sinc}\,[W(t - t_d)] \end{aligned} \tag{3.58b}$$



(a) (b)

Figure 3.28 Ideal low-pass filter frequency response and its impulse response.

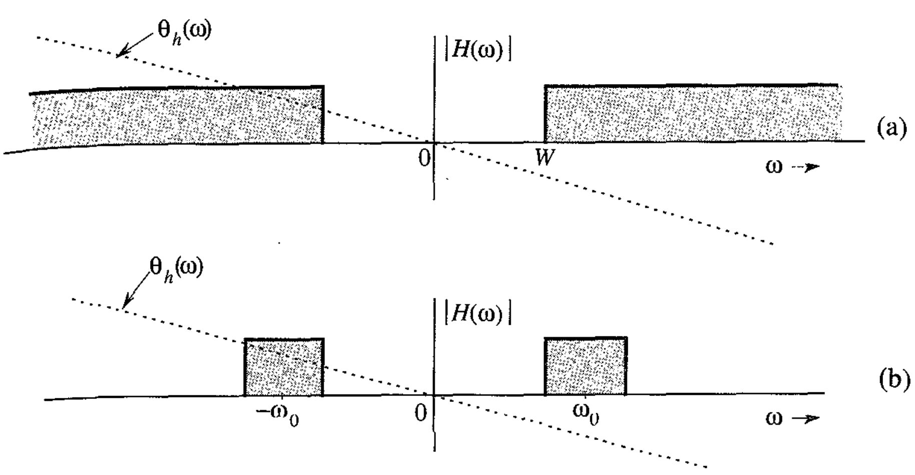


Figure 3.29 Ideal high-pass and bandpass filter frequency responses.

Recall that $h(t)$ is the system response to impulse input $\delta(t)$, which is applied at $t = 0$. Figure 3.28b shows a curious fact: the response $h(t)$ begins even before the input is applied (at $t = 0$). Clearly, the filter is noncausal and therefore physically unrealizable. Similarly, one can show that other ideal filters (such as the ideal high-pass or the ideal bandpass filters shown in Fig. 3.29) are also physically unrealizable.

For a physically realizable system, $h(t)$ must be causal; that is,

$$h(t) = 0 \qquad \text{for } t < 0$$

In the frequency domain, this condition is equivalent to the well-known **Paley-Wiener criterion**, which states that the necessary and sufficient condition for the amplitude response $|H(\omega)|$ to be realizable is*

$$\int_{-\infty}^{\infty} \frac{|\ln |H(\omega)||}{1+\omega^2} d\omega < \infty \tag{3.59}$$

If $H(\omega)$ does not satisfy this condition, it is unrealizable. Note that if $|H(\omega)| = 0$ over any finite band, $|\ln|H(\omega)|| = \infty$ over that band, and the condition (3.59) is violated. If, however, $H(\omega) = 0$ at a single frequency (or a set of discrete frequencies), the integral in Eq. (3.59) may still be finite even though the integrand is infinite. Therefore, for a physically realizable system, $H(\omega)$ may be zero at some discrete frequencies, but it cannot be zero over any finite band. According to this criterion, ideal filter characteristics (Figs. 3.28 and 3.29) are clearly unrealizable.

The impulse response $h(t)$ in Fig. 3.28 is not realizable. One practical approach to filter design is to cut off the tail of $h(t)$ for $t < 0$. The resulting causal impulse response $\widehat{h}(t)$, where

$$\widehat{h}(t) = h(t)u(t)$$

is physically realizable because it is causal (Fig. 3.30). If t_d is sufficiently large, $\widehat{h}(t)$ will be a close approximation of $h(t)$, and the resulting filter $\widehat{H}(\omega)$ will be a good approximation of an ideal filter. This close realization of the ideal filter is achieved because of the increased value

---

* $|H(\omega)|$ is assumed to be square integrable, that is,

$$\int_{-\infty}^{\infty} |H(\omega)|^2 \, d\omega$$

is finite. Note that the Paley-Wiener criterion is a criterion for the realizability of the amplitude response $|H(\omega)|$.

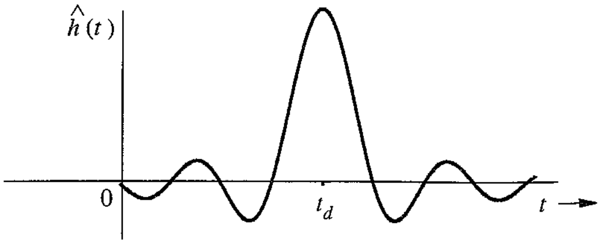


Figure 3.30 Approximate realization of an ideal low-pass filter by truncating its impulse response.

of time delay t_d. This means that the price of close realization is higher delay in the output; this is often true of noncausal systems. Of course, theoretically a delay $t_d = \infty$ is needed to realize the ideal characteristics. But a glance at Fig. 3.28b shows that a delay t_d of three or four times π/W will make $\hat{h}(t)$ a reasonably close version of $h(t - t_d)$. For instance, an audio filter is required to handle frequencies of up to 20 kHz. In this case a t_d of about 10^{-4} (0.1 ms) would be a reasonable choice. The truncation operation [cutting the tail of $h(t)$ to make it causal], however, creates some unsuspected problems of spectral spread and leakage, and can be partly corrected by truncating $h(t)$ gradually (rather than abruptly) using a tapered window function.[5]

In practice, we can realize a variety of filter characteristics to approach ideal characteristics. Practical (realizable) filter characteristics are gradual, without jump discontinuities in the amplitude response $|H(\omega)|$. The well-known Butterworth filters, for example, have amplitude response

$$|H(\omega)| = \frac{1}{\sqrt{1 + (\omega/2\pi B)^{2n}}}$$

These characteristics are shown in Fig. 3.31 for several values of n (the order of the filter). Note that the amplitude response approaches an ideal low-pass behavior as $n \rightarrow \infty$.

The half-power bandwidth of a filter is defined as the bandwidth over which the amplitude response $|H(\omega)|$ remains constant within variations of 3 dB (or a ratio of $1/\sqrt{2}$, that is, 0.707. Figure 3.31 shows that for all n, the Butterworth filter (half-power) bandwidth is B Hz. The

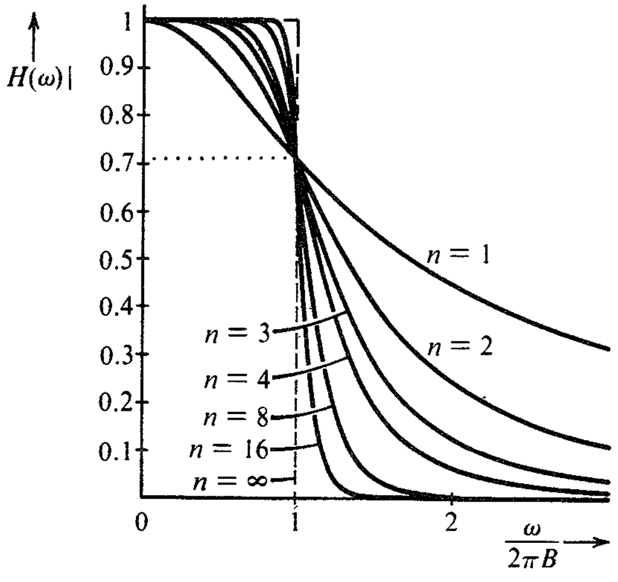


Figure 3.31 Butterworth filter characteristic.

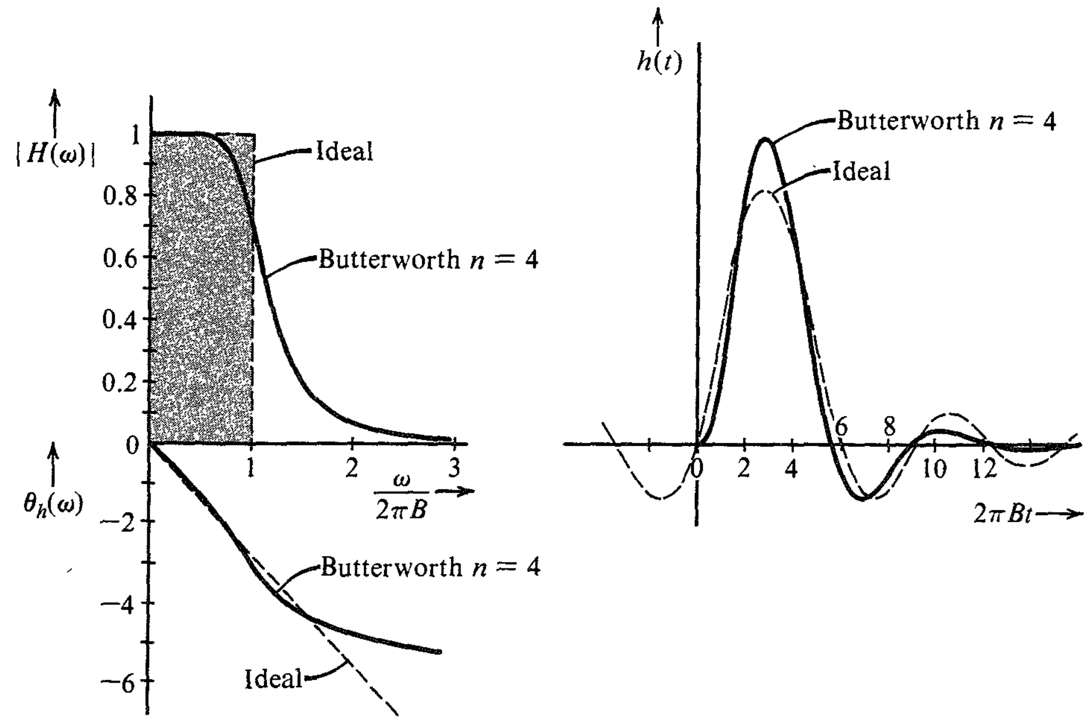


Figure 3.32 Comparison of Butterworth filter ($n = 4$) with an ideal filter.

half-power bandwidth of a low-pass filter is also called the **cutoff frequency**. Figure 3.32 shows $|H(\omega)|$, $\theta_h(\omega)$, and $h(t)$ for the case of $n = 4$.

It should be remembered that the magnitude $|H(\omega)|$ and the phase $\theta_h(\omega)$ of a system are interdependent; that is, we cannot choose $|H(\omega)|$ and $\theta_h(\omega)$ independently as we please. A certain trade-off exists between ideal magnitude and ideal phase characteristics. If we try to perfect $|H(\omega)|$ more, $\theta_h(\omega)$ deviates more from the ideal, and vice versa. As $n \to \infty$, the amplitude response approaches ideal, but the corresponding phase response is badly distorted in the vicinity of the cutoff frequency B Hz.

Digital Filters

Analog signals can also be processed by digital means (A/D conversion). This involves sampling, quantizing, and coding. The resulting digital signal can be processed by a small, special-purpose digital computer designed to convert the input sequence into a desired output sequence. The output sequence is converted back into the desired analog signal. A special algorithm of the processing digital computer can be used to achieve a given signal operation (e.g., low-pass, bandpass, or high-pass filtering).

Digital processing of analog signals has several advantages. A small, special-purpose computer can be time-shared for several uses, and the cost of digital implementation is often considerably lower than that of its analog counterpart. The accuracy of a digital filter is dependent only on the computer word length, the quantizing interval, and the sampling rate (aliasing error). Digital filters employ simple elements, such as adders, multipliers, shifters, and delay elements, rather than RLC components and operational amplifiers. As a result, they are generally unaffected by such factors as component accuracy, temperature stability, long-term drift, and so on, that afflict analog filter circuits. Also, many of the circuit restrictions imposed by physical limitations of analog devices can be removed, or at least circumvented, in a digital processor. Moreover, filters of a high order can be realized easily. Finally, digital filters can be modified simply by changing the algorithm of the computer, in contrast to an analog system, which may have to be physically rebuilt.

The subject of digital filtering is somewhat beyond our scope in this course. Several excellent books are available on the subject.[3]

3.6 SIGNAL DISTORTION OVER A COMMUNICATION CHANNEL

A signal transmitted over a channel is distorted because of various channel imperfections. The nature of signal distortion will now be studied.

Linear Distortion

We shall first consider linear time-invariant channels. Signal distortion can be caused over such a channel by nonideal characteristics of either the magnitude, the phase, or both. We can identify the effects these nonidealities will have on a pulse $g(t)$ transmitted through such a channel. Let the pulse exist over the interval (a, b) and be zero outside this interval. We recall the discussion in Sec. 3.1.1 about the marvelous balance of the Fourier spectrum. The components of the Fourier spectrum of the pulse have such a perfect and delicate balance of magnitudes and phases that they add up precisely to the pulse $g(t)$ over the interval (a, b) and to zero outside this interval. The transmission of $g(t)$ through an ideal channel that satisfies the conditions of distortionless transmission also leaves this balance undisturbed, because a distortionless channel multiplies each component by the same factor and delays each component by the same amount of time. Now, if the amplitude response of the channel is not ideal [that is, $|H(\omega)|$ is not equal to a constant], this delicate balance will be disturbed, and the sum of all the components cannot be zero outside the interval (a, b). In short, the pulse will spread out (see the following example). The same thing happens if the channel phase characteristic is not ideal, that is, $\theta_h(\omega) \neq -\omega t_d$. Thus, spreading, or **dispersion**, of the pulse will occur if either the amplitude response or the phase response, or both, are nonideal.

This type of distortion is undesirable in a TDM system, because pulse spreading causes interference with a neighboring pulse and consequently with a neighboring channel (crosstalk). For an FDM system, this type of distortion causes distortion (dispersion) in each multiplexed signal, but no interference occurs with a neighboring channel. This is because in FDM, each of the multiplexed signals occupies a band not occupied by any other signal. The amplitude and phase nonidealities of a channel will distort the spectrum of each signal, but because they are all nonoverlapping, no interference occurs among them.

EXAMPLE 3.17 A low-pass filter (Fig. 3.33a) transfer function $H(\omega)$ is given by

$$H(\omega) = \begin{cases} (1 + k\cos T\omega)e^{-j\omega t_d} & \omega| < 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \tag{3.60}$$

A pulse $g(t)$ band-limited to B Hz (Fig. 3.33b) is applied at the input of this filter. Find the output $y(t)$.

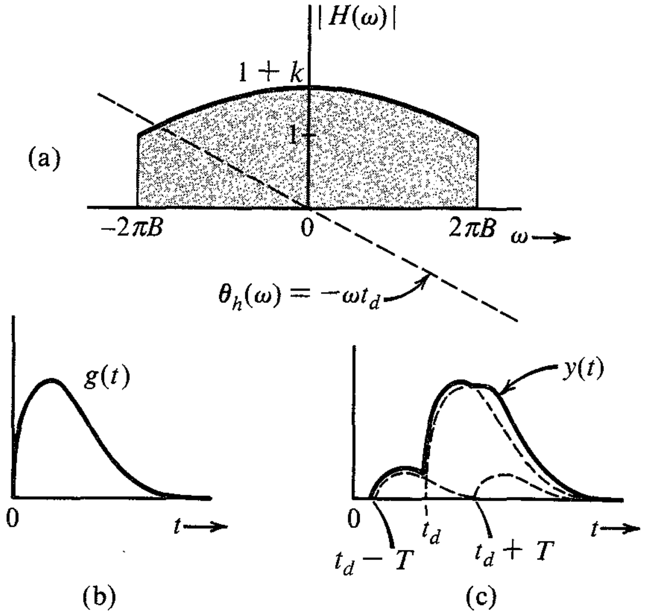


Figure 3.33 Pulse is dispersed when it passes through a system that is not distortionless.

This filter has ideal phase and nonideal magnitude characteristics. Because $g(t) \leftrightarrow G(\omega)$, $y(t) \leftrightarrow Y(\omega)$ and

$$\begin{aligned} Y(\omega) &= G(\omega)H(\omega) \\ &= G(\omega)(1 + k\cos T\omega)e^{-j\omega t_d} \\ &= G(\omega)e^{-j\omega t_d} + k[G(\omega)\cos T\omega]e^{-j\omega t_d} \end{aligned} \tag{3.61}$$

Using the time-shifting property and Eqs. (3.30a) and (3.32), we have

$$y(t) = g(t - t_d) + \frac{k}{2}[g(t - t_d - T) + g(t - t_d + T)] \tag{3.62}$$

The output is actually $g(t) + (k/2)[g(t - T) + g(t + T)]$ delayed by t_d. It consists of $g(t)$ and its echoes shifted by $\pm t_d$. The dispersion of the pulse caused by its echoes is evident from Fig. 3.33c. Ideal amplitude but nonideal phase response of $H(\omega)$ has a similar effect (see Prob. 3.6-1).

Distortion Caused by Channel Nonlinearities

Until now we considered the channel to be linear. This approximation is valid only for small signals. For large amplitudes, nonlinearities cannot be ignored. A general discussion of nonlinear systems is beyond our scope. Here we shall consider a simple case of a memoriless nonlinear channel where the input g and the output y are related by some nonlinear equation,

$$y = f(g)$$

The right-hand side of this equation can be expanded in a McLaurin's series as

$$y(t) = a_0 + a_1 g(t) + a_2 g^2(t) + a_3 g^3(t) + \cdots + a_k g^k(t) + \cdots$$

Recall the result in Sec. 3.3.6 (convolution) that if the bandwidth of $g(t)$ is B Hz, then the bandwidth of $g^k(t)$ is kB Hz.. Hence, the bandwidth of $y(t)$ is kB Hz. Consequently, the output spectrum spreads well beyond the input spectrum, and the output signal contains new frequency components not contained in the input signal. In broadcast communication, we need to amplify signals at very high power levels, where high-efficiency amplifiers (class C) are desirable. Unfortunately, these amplifiers are nonlinear, and their use to amplify signals causes distortion. This is one of the serious problems in AM signals. However, FM signals are not affected by nonlinear distortion, as shown in Chapter 5. If a signal is transmitted over a nonlinear channel, the nonlinearity not only distorts the signal, but also causes interference with other signals on the channel because of its spectral dispersion (spreading). The spectral dispersion will cause a serious interference problem in FDM systems (but not in TDM systems).

EXAMPLE 3.18 The input $x(t)$ and the output $y(t)$ of a certain nonlinear channel are related as

$$y(t) = x(t) + 0.001x^2(t)$$

Find the output signal $y(t)$ and its spectrum $Y(\omega)$ if the input signal is $x(t) = (1000/\pi)$ sinc $(1000t)$. Verify that the bandwidth of the output signal is twice that of the input signal. This is the result of signal squaring. Can the signal $x(t)$ be recovered (without distortion) from the output $y(t)$?

Since

$$x(t) = \frac{1000}{\pi} \text{ sinc } (1000t)$$

$$X(\omega) = \text{rect } \left(\frac{\omega}{2000}\right)$$

We have

$$y(t) = x(t) + 0.001x^2(t) = \frac{1000}{\pi} \text{ sinc } (1000t) + \frac{1000}{\pi^2} \text{ sinc}^2(1000t)$$

$$Y(\omega) = \text{rect } \left(\frac{\omega}{2000}\right) + 0.316\, \Delta \left(\frac{\omega}{4000}\right)$$

Observe that 0.316 $\text{sinc}^2(1000t)$ is the unwanted (distortion) term in the received signal. Figure 3.34a shows the input (desired) signal spectrum $X(\omega)$; Fig. 3.34b shows the spectrum of the undesired (distortion) term; and Fig. 3.34c shows the received signal spectrum $Y(\omega)$. We make the following observations:

1. The bandwidth of the received signal $y(t)$ is twice that of the input signal $x(t)$ (because of signal squaring).
2. The received signal contains the input signal $x(t)$ plus an unwanted signal $(1000/\pi)$ $\text{sinc}^2(1000t)$. The spectra of these two signals are shown in Fig. 3.34a and b. Figure 3.34c shows $Y(\omega)$, the spectrum of the received signal. Note that the desired signal and the distortion signal spectra overlap, and it is impossible to recover the signal $x(t)$ from the received signal $y(t)$ without some distortion.
3. We can reduce the distortion by passing the received signal through a low-pass filter of bandwidth 1000 rad/s. The spectrum of the output of this filter is shown in Fig. 3.34d.

Observe that the output of this filter is the desired input signal $x(t)$ with some residual distortion.

4. We have an additional problem of interference with other signals if the input signal $x(t)$ is frequency-division multiplexed along with several other signals on this channel. This means that several signals occupying nonoverlapping frequency bands are transmitted simultaneously on the same channel. Spreading of the spectrum $X(\omega)$ outside its original band of 1000 rad/s will interfere with the signal in the band of 1000 to 2000 rad/s. Thus, in addition to the distortion of $x(t)$, we also have an interference with the neighboring band.
5. If $x(t)$ were a digital signal consisting of a pulse train, each pulse would be distorted, but there would be no interference with the neighboring pulses. Moreover even with distorted pulses, data can be received without loss because digital communication can withstand considerable pulse distortion without loss of information. Thus, if this channel were used to transmit a TDM signal consisting of two interleaved pulse trains, the data in the two trains would be recovered at the receiver.

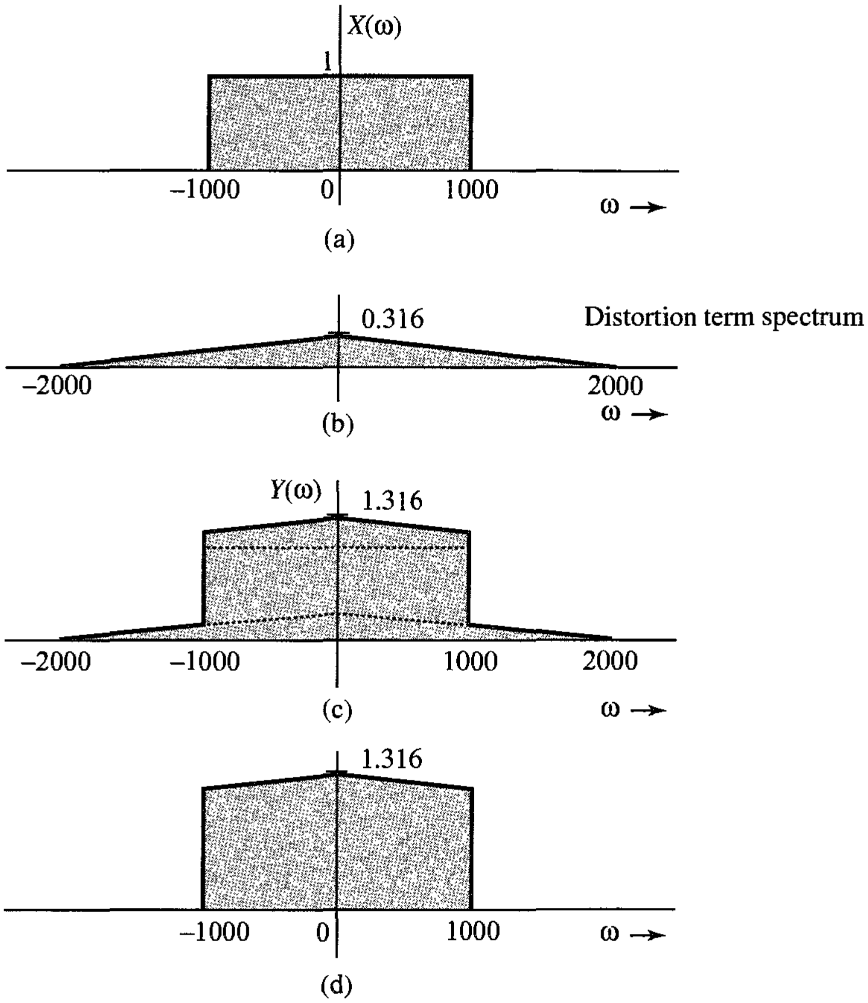


Figure 3.34 Signal distortion caused by nonlinear operation. (a) Desired (input) signal spectrum. (b) Spectrum of the unwanted signal (distortion) in the received signal. (c) Spectrum of the received signal. (d) Spectrum of the received signal after low-pass filtering.

Distortion Caused by Multipath Effects

A multipath transmission takes place when a transmitted signal arrives at the receiver by two or more paths of different delays. For example, if a signal is transmitted over a cable that has impedance irregularities (mismatching) along the path, the signal will arrive at the receiver in the form of a direct wave plus various reflections with various delays. In radio links, the signal can be received by direct path between the transmitting and the receiving antennas and also by reflections from other objects, such as hills, buildings, and so on. In long-distance radio links using the ionosphere, similar effects occur because of one-hop and multihop paths. In each of these cases, the transmission channel can be represented as several channels in parallel, each with a different relative attenuation and a different time delay. Let us consider the case of only two paths: one with a unity gain and a delay t_d, and the other with a gain α and a delay $t_d + \Delta t$, as shown in Fig. 3.35a. The transfer functions of the two paths are given by $e^{-j\omega t_d}$ and $\alpha e^{-j\omega(t_d+\Delta t)}$, respectively. The overall transfer function of such a channel is $H(\omega)$, given by

$$\begin{aligned}
H(\omega) &= e^{-j\omega t_d} + \alpha e^{-j\omega(t_d+\Delta t)} \\
&= e^{-j\omega t_d}(1 + \alpha e^{-j\omega\Delta t}) && \text{(3.63a)} \\
&= e^{-j\omega t_d}(1 + \alpha\cos\omega\Delta t - j\alpha\sin\omega\Delta t) \\
&= \underbrace{\sqrt{1+\alpha^2+2\alpha\cos\omega\Delta t}}_{|H(\omega)|}\; e^{-j\underbrace{\left[\omega t_d + \tan^{-1}\frac{\alpha\sin\omega\Delta t}{1+\alpha\cos\omega\Delta t}\right]}_{\theta_h(\omega)}} && \text{(3.63b)}
\end{aligned}$$

Both the magnitude and the phase characteristics of $H(\omega)$ are periodic in ω with a period of $2\pi/\Delta t$ (Fig. 3.35b). The multipath transmission, therefore, causes nonidealities in

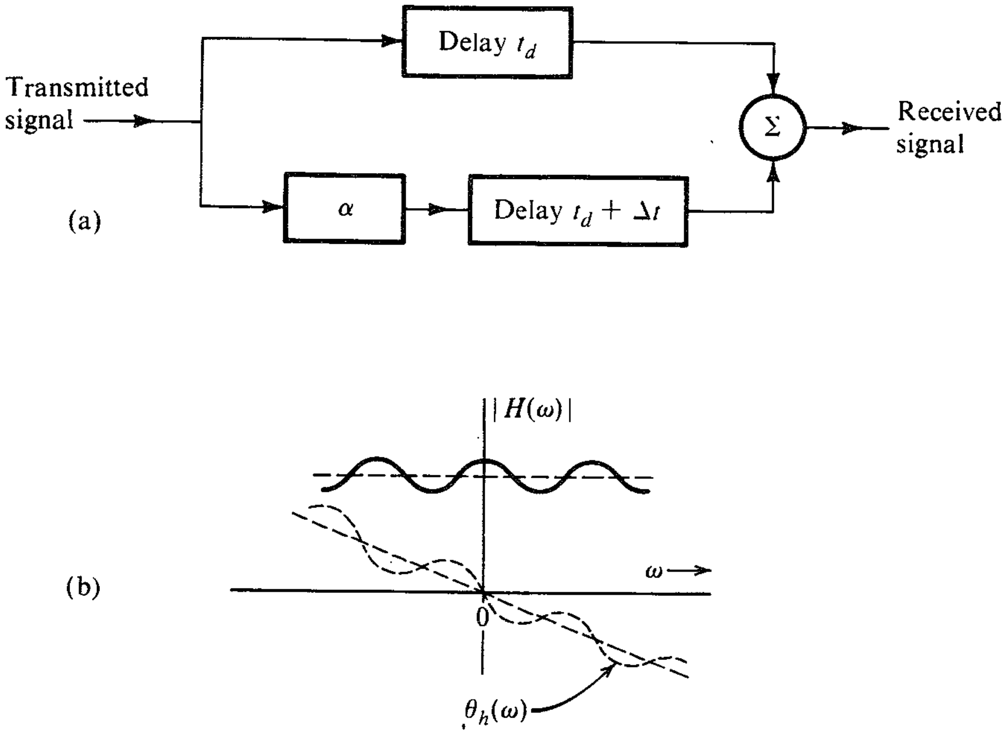


Figure 3.35 Multipath transmission.

the magnitude and the phase characteristics of the channel and will cause linear distortion (pulse dispersion), as discussed earlier. If, for instance, the gains of the two paths are very close, that is, $\alpha \approx 1$, the signals received by the two paths can very nearly cancel each other at certain frequencies, where their phases are π rad apart (signal annihilation by destructive interference). Equation (3.63b) shows that at frequencies where $\omega = n\pi/\Delta t$ (n odd), $\cos \omega\Delta t = -1$, and $|H(\omega)| \approx 0$ when $\alpha \approx 1$. These frequencies are the multipath null frequencies. At frequencies $\omega = n\pi/\Delta t$ (n even), the two signals interfere constructively to enhance the gain. Such channels cause **frequency-selective fading** of transmitted signals. Such distortion can be partly corrected by using the tapped delay-line equalizer, as shown in Prob. 3.6-2 These equalizers are useful in several applications in communications, discussed in Chapters 6 and 7.

Fading Channels

Thus far, the channel characteristics were assumed to be constant with time. In practice, we encounter channels whose transmission characteristics vary with time. These include troposcatter channels and channels using the ionosphere for radio reflection to achieve long-distance communication. The time variations of the channel properties arise because of semiperiodic and random changes in the propagation characteristics of the medium. The reflection properties of the ionosphere, for example, are related to meteorological conditions that change seasonally, daily, and even from hour to hour, much the same way as does the weather. Periods of sudden storms also occur. Hence, the effective channel transfer function varies semiperiodically and randomly, causing random attenuation of the signal. This phenomenon is known as **fading**. One way to reduce the effects of fading is to use **automatic gain control (AGC)**.[*]

Fading may be strongly frequency dependent where different frequency components are affected unequally. Such fading is known as frequency-selective fading and can cause serious problems in communication. Multipath propagation can cause frequency-selective fading.

3.7 SIGNAL ENERGY AND ENERGY SPECTRAL DENSITY

The energy E_g of a signal $g(t)$ is defined as the area under $|g(t)|^2$. We can also determine the signal energy from its Fourier transform $G(\omega)$ through Parseval's theorem.

Parseval's Theorem

Signal energy can be related to the signal spectrum $G(\omega)$ by substituting Eq. (3.8b) in Eq. (2.2):

$$E_g = \int_{-\infty}^{\infty} g(t)g^*(t)\,dt = \int_{-\infty}^{\infty} g(t)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} G^*(\omega)e^{-j\omega t}\,d\omega\right]dt$$

Here, we used the fact that $g^*(t)$, being the conjugate of $g(t)$, can be expressed as the conjugate of the right-hand side of Eq. (3.8b). Now, interchanging the order of integration yields

$$\begin{aligned} E_g &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G^*(\omega)\left[\int_{-\infty}^{\infty} g(t)e^{-j\omega t}\,dt\right]d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)G^*(\omega)\,d\omega \end{aligned}$$

---

[*] AGC will also suppress slow variations of the original signal.

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(\omega)|^2 \, d\omega \tag{3.64}$$

This is the statement of the well-known Parseval's theorem. A similar result was obtained for a periodic signal and its Fourier series in Eq. (2.59). This result allows us to determine the signal energy from either the time-domain specification $g(t)$ or the frequency-domain specification $G(\omega)$ of the same signal.

EXAMPLE 3.19 Verify Parseval's theorem for the signal $g(t) = e^{-at}u(t)$ $(a > 0)$.

We have

$$E_g = \int_{-\infty}^{\infty} g^2(t)\, dt = \int_0^{\infty} e^{-2at}\, dt = \frac{1}{2a} \tag{3.65}$$

We now determine E_g from the signal spectrum $G(\omega)$ given by

$$G(\omega) = \frac{1}{j\omega + a}$$

and from Eq. (3.64),

$$E_g = \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(\omega)|^2 \, d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{\omega^2 + a^2}\, d\omega = \frac{1}{2\pi a} \tan^{-1}\frac{\omega}{a}\bigg|_{-\infty}^{\infty} = \frac{1}{2a}$$

which verifies Parseval's theorem.

Energy Spectral Density (ESD)

Equation (3.64) can be interpreted to mean that the energy of a signal $g(t)$ is the result of energies contributed by all the spectral components of the signal $g(t)$. The contribution of a spectral component of frequency ω is proportional to $|G(\omega)|^2$. To elaborate this further, consider a signal $g(t)$ applied at the input of an ideal bandpass filter, whose transfer function $H(\omega)$ is shown in Fig. 3.36a. This filter suppresses all frequencies except a narrow band $\Delta\omega$ $(\Delta\omega \to 0)$ centered at a frequency ω_0 (Fig. 3.36b). If the filter output is $y(t)$, then its Fourier transform $Y(\omega) = G(\omega)H(\omega)$, and E_y, the energy of the output $y(t)$, is

$$E_y = \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(\omega)H(\omega)|^2 \, d\omega \tag{3.66}$$

Because $H(\omega)$=1 over the passband $\Delta\omega$, and zero everywhere else, the intergral on the right-hand side is the sum of the two shaded areas in Fig. 3.36b, and we have (for $\Delta\omega \to 0$)

$$E_y = 2\,\frac{1}{2\pi}\,|G(\omega_0)|^2 \, d\omega = 2|G(\omega_0)|^2 \, df$$

Thus, $2|G(\omega)|^2\, df$ is the energy contributed by the spectral components within the two narrow bands, each of width Δf Hz, centered at $\pm\omega_0$. Therefore, we can interpret $|G(\omega)|^2$ as the energy per unit bandwidth (in hertz) of the spectral components of $g(t)$ centered at frequency ω. In other words, $|G(\omega)|^2$ is the energy spectral density (per unit bandwidth in hertz) of $g(t)$. Actually the energy contributed per unit bandwidth is $2|G(\omega)|^2$ because both the positive and the negative frequency components combine to form the components in the band Δf. However, for the sake of convenience we consider the positive and negative frequency components being

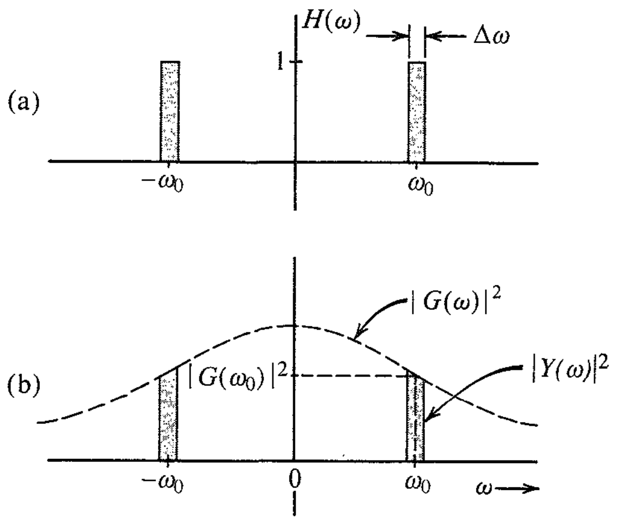


Figure 3.36 Interpretation of the energy spectral density of a signal.

independent. Some authors *do* define $2|G(\omega)|^2$ as the energy spectral density. The **energy spectral density (ESD)** $\Psi_g(t)$ is thus defined as

$$\Psi_g(\omega) = |G(\omega)|^2 \tag{3.67}$$

and Eq. (3.64) can be expressed as

$$E_g = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_g(\omega)\, d\omega = \int_{-\infty}^{\infty} \Psi_g(\omega)\, df \tag{3.68}$$

From the results in Example 3.19, the ESD of the signal $g(t) = e^{-at}u(t)$ is

$$\Psi_g(\omega) = |G(\omega)|^2 = \frac{1}{\omega^2 + a^2}$$

Essential Bandwidth of a Signal

The spectra of most signals extend to infinity. However, because the energy of a practical signal is finite, the signal spectrum must approach 0 as $\omega \to \infty$. Most of the signal energy is contained within a certain band of B Hz, and the energy content of the components of frequencies greater than B Hz is negligible. We can therefore suppress the signal spectrum beyond B Hz with little effect on the signal shape and energy. The bandwidth B is called the **essential bandwidth** of the signal. The criterion for selecting B depends on the error tolerance in a particular application. We may, for instance, select B to be that band which contains 95% of the signal energy[*]. This figure may be higher or lower than 95%, depending on the precision needed. Using such a criterion, we can determine the essential bandwidth of a signal. Suppression of all the spectral components of $g(t)$ beyond the essential bandwidth results in a signal $\hat{g}(t)$, which is a close approximation of $g(t)$.[†] If we use the 95% criterion for the essential bandwidth, the energy of the error (the difference) $g(t) - \hat{g}(t)$ is 5% of E_g. The following example demonstrates the bandwidth estimation procedure.

---

[*] Essential bandwidth for a low-pass signal may also be defined as a frequency at which the value of the amplitude spectrum is a small fraction (about 5 to 10%) of its peak value. In Example 3.19, the peak of $|G(\omega)|$ is $1/a$, and it occurs at $\omega = 0$.

[†] In practice the truncation is performed gradually using tapered windows in order to avoid excessive spectral leakage resulting from the abrupt truncation.[5]

EXAMPLE 3.20 Estimate the essential bandwidth W rad/s of the signal $e^{-at}u(t)$ if the essential band is required to contain 95% of the signal energy.

In this case,

$$G(\omega) = \frac{1}{j\omega + a}$$

and the ESD is

$$|G(\omega)^2| = \frac{1}{\omega^2 + a^2}$$

This ESD is shown in Fig. 3.37. Moreover, the signal energy E_g is $1/2\pi$ times the area under this ESD, which has already been found to be $1/2a$. Let W rad/s be the essential bandwidth, which contains 95% of the total signal energy E_g. This means $1/2\pi$ times the shaded area in Fig. 3.37 is $0.95/2a$, that is,

$$\begin{aligned}\frac{0.95}{2a} &= \frac{1}{2\pi}\int_{-W}^{W}\frac{d\omega}{\omega^2 + a^2} \\ &= \frac{1}{2\pi a}\tan^{-1}\frac{\omega}{a}\bigg|_{-W}^{W} = \frac{1}{\pi a}\tan^{-1}\frac{W}{a}\end{aligned}$$

or

$$\frac{0.95\pi}{2} = \tan^{-1}\frac{W}{a} \Longrightarrow W = 12.706a \text{ rad/s}$$

This means that the spectral components of $g(t)$ in the band from 0 (dc) to 12.706 rad/s (2.02 Hz) contribute 95% of the total signal energy; all the remaining spectral components (in the band from 12.706 rad/s to ∞) contribute only 5% of the signal energy.*

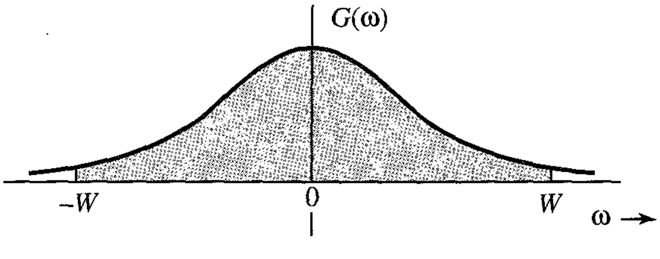


Figure 3.37 Estimating the essential bandwidth of a signal.

EXAMPLE 3.21 Estimate the essential bandwidth of a rectangular pulse $g(t) = \text{rect}(t/T)$ (Fig. 3.38a), where the essential bandwidth is to contain at least 90% of the pulse energy.

For this pulse, the energy E_g is

$$E_g = \int_{-\infty}^{\infty} g^2(t)\,dt = \int_{-T/2}^{T/2} dt = T$$

---

[*] Note that although the ESD exists over the band $-\infty$ to ∞, the trigonometric spectrum exists only over the band 0 to ∞. The spectrum range $-\infty$ to ∞ applies to the exponential spectrum. In practice, whenever we talk about a bandwidth, we mean it in the trigonometric sense. Hence, the essential band is from 0 to W, not $-W$ to W.

Also because

$$\operatorname{rect}\left(\frac{t}{T}\right) \Longleftrightarrow T \operatorname{sinc}\left(\frac{\omega T}{2}\right)$$

the ESD for this pulse is

$$\Psi_g(\omega) = |G(\omega)|^2 = T^2 \operatorname{sinc}^2\left(\frac{\omega T}{2}\right)$$

This ESD is shown in Fig. 3.38b as a function of ωT as well as fT, where f is the frequency in hertz. The energy E_W within the band from 0 to W rad/s is given by

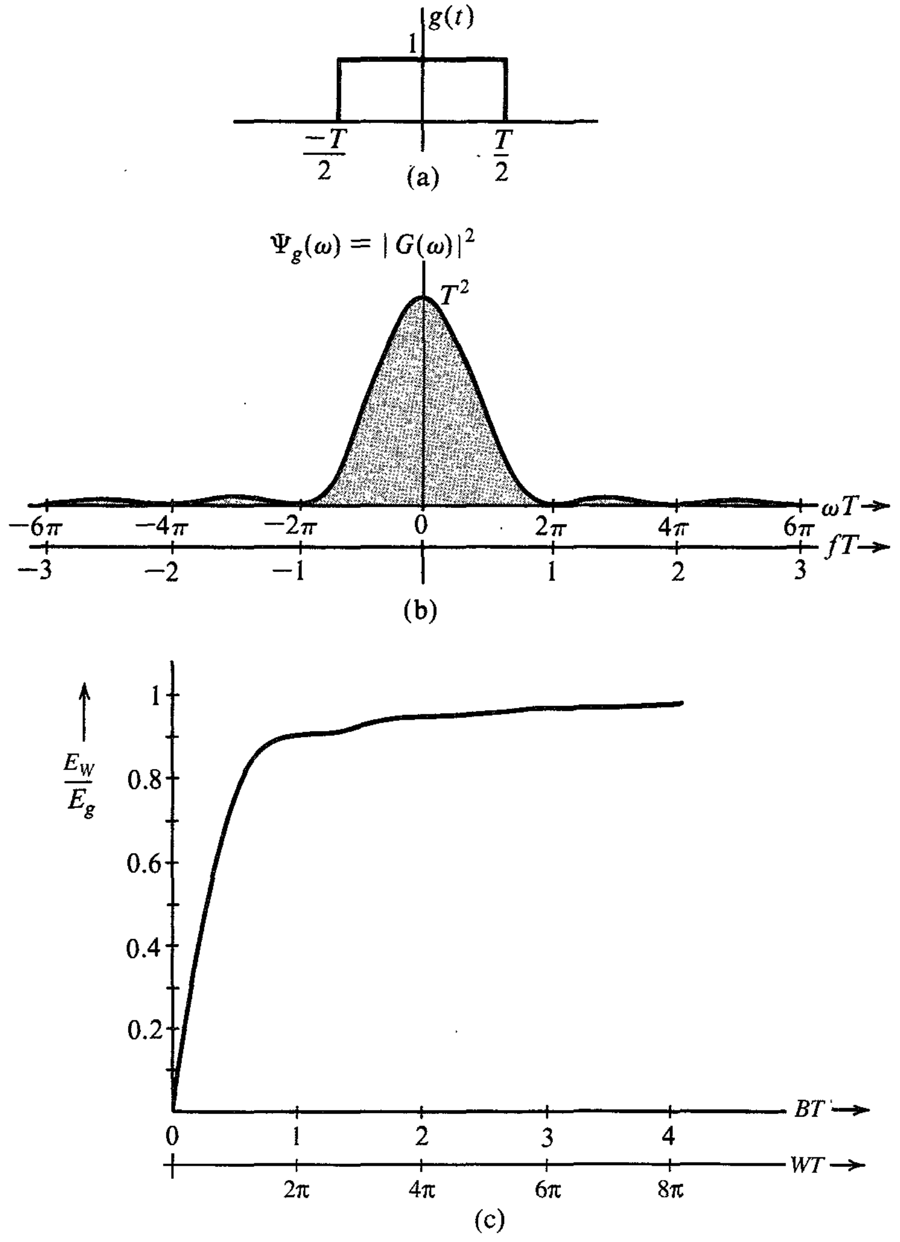


Figure 3.38 Gate function and its energy spectral density.

$$E_W = \frac{1}{2\pi}\int_{-W}^{W} T^2 \operatorname{sinc}^2\left(\frac{\omega T}{2}\right) d\omega$$

Setting $\omega T = x$ in this integral so that $d\omega = (1/T)\ dx$, we obtain

$$E_W = \frac{T}{\pi}\int_{0}^{WT} \operatorname{sinc}^2\left(\frac{x}{2}\right) dx$$

Also because $E_g = T$, we have

$$\frac{E_W}{E_g} = \frac{1}{\pi}\int_{0}^{WT} \operatorname{sinc}^2\left(\frac{x}{2}\right) dx$$

The integral on the right-hand side is numerically computed, and the plot of E_W/E_g vs. WT is shown in Fig. 3.38c. Note that 90.28% of the total energy of the pulse $g(t)$ is contained within the band $W = 2\pi/T$ rad/s or $B = 1/T$ Hz. Therefore, using the 90% criterion, the bandwidth of a rectangular pulse of width T seconds is $1/T$ Hz. A similar result was obtained from Example 3.2.

Energy of Modulated Signals

We have seen that modulation shifts the signal spectrum $G(\omega)$ to the left and right by ω_0. We now show that a similar thing happens to the ESD of the modulated signal.

Let $g(t)$ be a baseband signal band-limited to B Hz. The amplitude-modulated signal $\varphi(t)$ is

$$\varphi(t) = g(t)\ \cos\omega_0 t$$

and the spectrum (Fourier transform) of $\varphi(t)$ is

$$\Phi(\omega) = \frac{1}{2}[G(\omega+\omega_0) + G(\omega-\omega_0)]$$

The ESD of the modulated signal $\varphi(t)$ is $|\Phi(\omega)|^2$, that is,

$$\Psi_\varphi(\omega) = \frac{1}{4}|G(\omega+\omega_0) + G(\omega-\omega_0)|^2$$

If $\omega_0 \geq 2\pi B$, then $G(\omega+\omega_0)$ and $G(\omega-\omega_0)$ are nonoverlapping (see Fig. 3.39), and

$$\Psi_\varphi(\omega) = \frac{1}{4}\Big[|G(\omega+\omega_0)|^2 + |G(\omega-\omega_0)|^2\Big] \tag{3.69a}$$

$$= \frac{1}{4}\Big[\Psi_g(\omega+\omega_0) + \Psi_g(\omega-\omega_0)\Big] \tag{3.69b}$$

The ESDs of both $g(t)$ and the modulated signal $\varphi(t)$ are shown in Fig. 3.39. It is clear that modulation shifts the ESD of $g(t)$ by $\pm\omega_0$. Observe that the area under $\Psi_\varphi(\omega)$ is half the area under $\Psi_g(\omega)$. Because the energy of a signal is proportional to the area under its ESD, it follows that the energy of $\varphi(t)$ is half the energy of $g(t)$, that is,

$$E_\varphi = \frac{1}{2}E_g \qquad \omega_0 \geq 2\pi B \tag{3.70}$$

It may seem surprising that a signal $\varphi(t)$, which appears so energetic compared to $g(t)$, should have only half the energy of $g(t)$. Appearances are deceiving, as usual. The energy of a signal

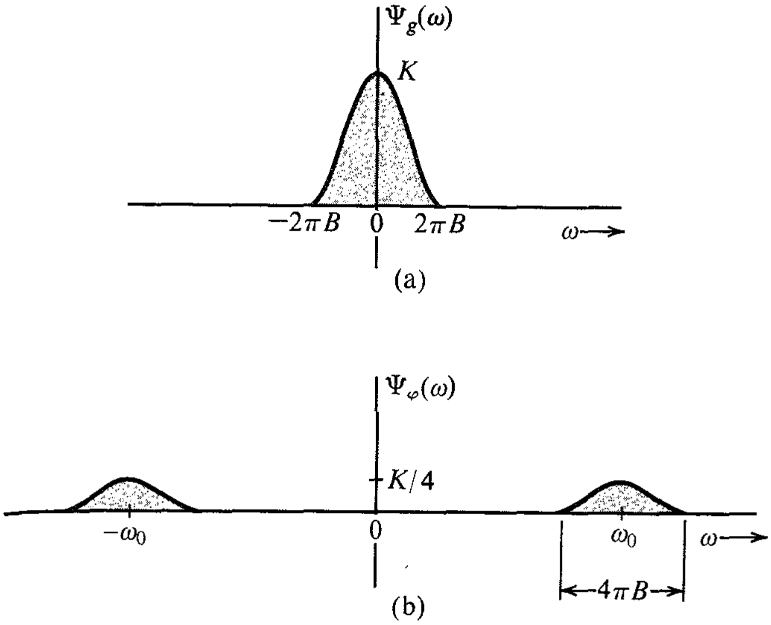


Figure 3.39 Energy spectral densities of modulating and modulated signals.

is proportional to the square of its amplitude, and higher amplitudes contribute more energy. Signal $g(t)$ remains at higher amplitude levels most of the time. On the other hand, $\varphi(t)$, because of the factor $\cos \omega_0 t$, dips to zero amplitude levels many times, which reduces its energy.

Time Autocorrelation Function and the Energy Spectral Density

In Chapter 2, we showed that a good measure of comparing two signals $g(t)$ and $z(t)$ is the correlation function $\psi_{gz}(\tau)$ defined in Eq. (2.49). We also defined the correlation of a signal $g(t)$ with itself [the autocorrelation function $\psi_g(\tau)$] in Eq. (2.50). For a real signal $g(t)$, the autocorrelation function $\psi_g(\tau)$ is given by*

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t+\tau)\, dt \tag{3.71a}$$

Setting $x = t + \tau$ in Eq. (3.71a) yields

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(x)g(x-\tau)\, dx$$

In this equation, x is a dummy variable and could be replaced by t. Thus,

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t \pm \tau)\, dt \tag{3.71b}$$

This shows that for a real $g(t)$, the autocorrelation function is an even function of τ, that is,

$$\psi_g(\tau) = \psi_g(-\tau) \tag{3.72}$$

---

* For a complex signal $g(t)$, we define

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g^*(t)g(t+\tau)\, dt \tag{3.71n}$$

We now show that the ESD $\Psi_g(\omega) = |G(\omega)|^2$ is the Fourier transform of the autocorrelation function $\psi_g(\tau)$. Although the result is proved here for real signals, it is valid for complex signals also. Note that the autocorrelation function is a function of τ, not t. Hence, its Fourier transform is $\int \psi(\tau)e^{-j\omega\tau}\,d\tau$. Thus,

$$\begin{aligned}\mathcal{F}[\psi_g(\tau)] &= \int_{-\infty}^{\infty} e^{-j\omega\tau}\left[\int_{-\infty}^{\infty} g(t)g(t+\tau)\,dt\right] d\tau \\ &= \int_{-\infty}^{\infty} g(t)\left[\int_{-\infty}^{\infty} g(\tau+t)e^{-j\omega\tau}d\tau\right] dt\end{aligned}$$

The inner integral is the Fourier transform of $g(\tau+t)$, which is $g(\tau)$ left-shifted by t. Hence, it is given by [time-shifting property in Eq. (3.30)] $G(\omega)e^{j\omega t}$. Therefore,

$$\mathcal{F}[\psi_g(\tau)] = G(\omega)\int_{-\infty}^{\infty} g(t)e^{j\omega t}dt = G(\omega)G(-\omega) = |G(\omega)|^2$$

This shows that

$$\psi_g(\tau) \Longleftrightarrow \Psi_g(\omega) = |G(\omega)|^2 \tag{3.73}$$

A careful observation of the operation of correlation shows close connection to convolution. Indeed, the autocorrelation function $\psi_g(\tau)$ is the convolution of $g(\tau)$ with $g(-\tau)$ because

$$g(\tau) * g(-\tau) = \int_{-\infty}^{\infty} g(x)g[-(\tau-x)]\,dx = \int_{-\infty}^{\infty} g(x)g(x-\tau)\,dx = \psi_g(\tau)$$

Application of the time convolution property [Eq. (3.43)] to this equation yields Eq. (3.73).

EXAMPLE 3.22 Find the time autocorrelation function of the signal $g(t) = e^{-at}u(t)$, and from it determine the ESD of $g(t)$.

In this case,

$$g(t) = e^{-at}u(t) \quad \text{and} \quad g(t-\tau) = e^{-a(t-\tau)}u(t-\tau)$$

Recall that $g(t-\tau)$ is $g(t)$ right-shifted by τ, as shown in Fig. 3.40a (for positive τ). The autocorrelation function $\psi_g(\tau)$ is given by the area under the product $g(t)g(t-\tau)$ [see Eq. (3.71b)]. Therefore,

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t-\tau)\,dt = e^{a\tau}\int_{\tau}^{\infty} e^{-2at}\,dt = \frac{1}{2a}e^{-a\tau}$$

This is valid for positive τ. We can perform a similar procedure for negative τ. However, we know that for a real $g(t)$, $\psi_g(\tau)$ is an even function of τ. Therefore,

$$\psi_g(\tau) = \frac{1}{2a}e^{-a|\tau|}$$

Figure 3.40b shows the autocorrelation function $\psi_g(\tau)$. The ESD $\Psi_g(\omega)$ is the Fourier transform of $\psi_g(\tau)$. From Table 3.1 (pair 3), it follows that

$$\Psi_g(\omega) = \frac{1}{\omega^2 + a^2}$$

which confirms the earlier result.

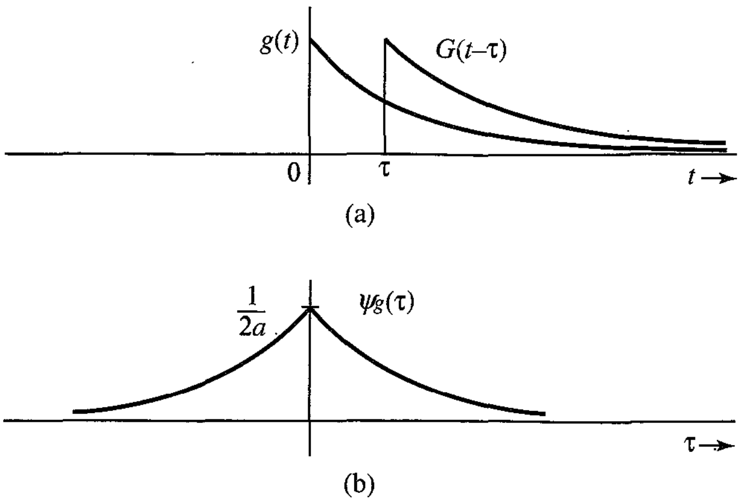


Figure 3.40 Computation of the time autocorrelation function.

ESD of the Input and the Output

If $g(t)$ and $y(t)$ are the input and the corresponding output of a linear time-invariant (LTI) system, then

$$Y(\omega) = H(\omega)G(\omega)$$

Therefore,

$$|Y(\omega)|^2 = |H(\omega)|^2|G(\omega)|^2$$

This shows that

$$\Psi_y(\omega) = |H(\omega)|^2\Psi_g(\omega) \tag{3.74}$$

Thus, the output signal ESD is $|H(\omega)|^2$ times the input signal ESD.

3.8 SIGNAL POWER AND POWER SPECTRAL DENSITY

For a power signal, a meaningful measure of its size is its power [defined in Eq. (2.4)] as the time average of the signal energy averaged over the infinite time interval. The power P_g of a real signal $g(t)$ is given by

$$P_g = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} g^2(t)\,dt \tag{3.75}$$

The signal power and the related concepts can be readily understood by defining a truncated signal $g_T(t)$ as

$$g_T(t) = \begin{cases} g(t) & |t| \le T/2 \\ 0 & |t| > T/2 \end{cases}$$

The truncated signal is shown in Fig. 3.41. The integral on the right-hand side of Eq. (3.75) is the energy of the truncated signal $g_T(t)$. Thus,

$$P_g = \lim_{T\to\infty} \frac{E_{g_T}}{T} \tag{3.76}$$

This equation serves as a link between power and energy. Understanding this relationship will be very helpful in understanding and relating all the power concepts to the energy concepts. Because the signal power is just the time average of energy, all the concepts and results of signal energy apply to signal power also if we modify the concepts properly by taking their time averages.

Power Spectral Density (PSD)

If the signal $g(t)$ is a power signal, then its power is finite, and the truncated signal $g_T(t)$ is an energy signal as long as T is finite. If $g_T(t) \Longleftrightarrow G_T(\omega)$, then from Parseval's theorem,

$$E_{g_T} = \int_{-\infty}^{\infty} g_T^2(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |G_T(\omega)|^2\,d\omega$$

Hence, P_g, the power of $g(t)$, is given by

$$P_g = \lim_{T\to\infty} \frac{E_{g_T}}{T} = \lim_{T\to\infty} \frac{1}{T}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} |G_T(\omega)|^2\,d\omega\right] \tag{3.77}$$

As T increases, the duration of $g_T(t)$ increases, and its energy E_{g_T} also increases proportionately. This means $|G_T(\omega)|^2$ also increases with T, and as $T \to \infty$, $|G_T(\omega)|^2$ also approaches ∞. However, $|G_T(\omega)|^2$ must approach ∞ at the same rate as T because for a power signal, the right-hand side of Eq. (3.77) must converge. This convergence permits us to interchange the order of the limiting process and integration in Eq. (3.77), and we have

$$P_g = \frac{1}{2\pi}\int_{-\infty}^{\infty} \lim_{T\to\infty} \frac{|G_T(\omega)|^2}{T}\,d\omega \tag{3.78}$$

We define the **power spectral density (PSD)** $S_g(\omega)$ as

$$S_g(\omega) = \lim_{T\to\infty} \frac{|G_T(\omega)|^2}{T} \tag{3.79}$$

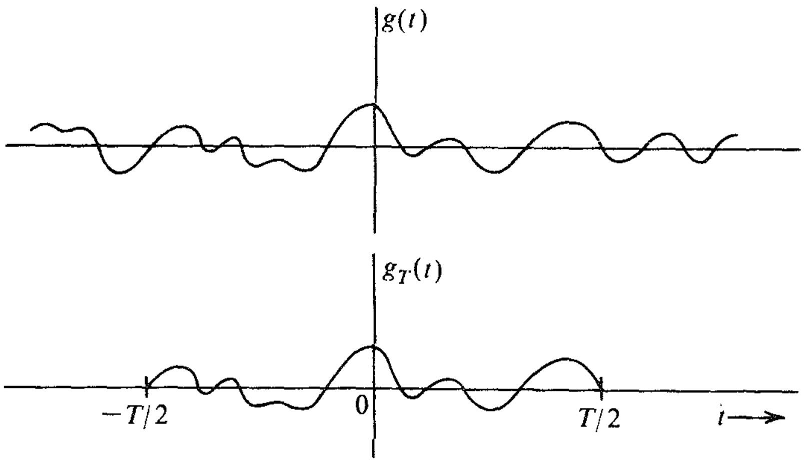


Figure 3.41 Limiting process in derivation of PSD.

Consequently,*

$$P_g = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_g(\omega)\,d\omega \tag{3.80a}$$

$$= \frac{1}{\pi}\int_{0}^{\infty} S_g(\omega)\,d\omega \tag{3.80b}$$

This result is parallel to the result [Eq. (3.68)] for energy signals. The power is $1/2\pi$ times the area under the PSD. Observe that the PSD is the time average of the ESD of $g_T(t)$ [Eq. (3.79)].

As is the case with ESD, the PSD is also a positive, real, and even function of ω. If $g(t)$ is a voltage signal, the units of PSD are volts squared per hertz. Equations (3.80) can be expressed in a more compact form using the variable f (in hertz) as

$$P_g = \int_{-\infty}^{\infty} S_g(\omega)\,df = 2\int_{0}^{\infty} S_g(\omega)\,df \tag{3.81}$$

Time-Autocorrelation Function of Power Signals

The (time) autocorrelation function $\mathcal{R}_g(\tau)$ of a real power signal $g(t)$ is defined as†

$$\mathcal{R}_g(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} g(t)g(t+\tau)\,dt \tag{3.82a}$$

Using the same argument as that used for energy signals [Eqs. (3.71b) and (3.72)], we can show that $\mathcal{R}_g(\tau)$ is an even function of τ. This means for a real $g(t)$

$$\mathcal{R}_g(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} g(t)g(t-\tau)\,dt \tag{3.82b}$$

and

$$\mathcal{R}_g(\tau) = \mathcal{R}_g(-\tau) \tag{3.83}$$

For energy signals, the ESD $\Psi_g(\omega)$ is the Fourier transform of the autocorrelation function $\psi_g(\tau)$. A similar result applies to power signals. We now show that for a power signal, the PSD $S_g(\omega)$ is the Fourier transform of the autocorrelation function $\mathcal{R}_g(\tau)$. From Eq. (3.82a) and Fig. 3.41,

$$\mathcal{R}_g(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} g_T(t)g_T(t+\tau)\,dt = \lim_{T\to\infty}\frac{\psi_{g_T}(\tau)}{T}$$

Recall that $\psi_{g_T}(\tau) \Longleftrightarrow |G_T(\omega)|^2$. Hence, the Fourier transform of the preceding equation yields

$$\mathcal{R}_g(\tau) \Longleftrightarrow \lim_{T\to\infty}\frac{|G_T(\omega)|^2}{T} = S_g(\omega) \tag{3.84}$$

---

[*] One should use caution in using a unilateral expression such as $P_g = 2(1/2\pi)\int_0^{\infty} S_g(\omega)\,d\omega$ when $S_g(\omega)$ contains an impulse at the origin (a dc component). The impulse part should not be multiplied by the factor 2.

[†] For a complex $g(t)$, we define

$$\mathcal{R}_g(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} g^*(t)g(t+\tau)\,dt \tag{3.82n}$$

Although we have proved these results for a real $g(t)$, Eqs. (3.79), (3.80), (3.81), and (3.84) are equally valid for a complex $g(t)$.

The concept and relationships for signal power are parallel to those for signal energy. This is brought out in Table 3.3.

Signal Power Is Its Mean Square Value

A glance at Eq. (3.75) shows that the signal power is the time average or mean of its squared value. In other words P_g is the mean square value of $g(t)$. We must remember, however, that this is a time mean, not a statistical mean (to be discussed in later chapters). Statistical means are denoted by overbars. Thus, the (statistical) mean square of a variable x is denoted by $\overline{x^2}$. To distinguish from this kind of mean, we shall use a wavy overline to denote a time average. Thus, the time mean square value of $g(t)$ will be denoted by $\widetilde{g^2(t)}$. The time averages are conventionally denoted by pointed brackets, such as $< g^2(t) >$. We shall, however, use the wavy overline notation because it is much easier to associate means with a bar on top rather than the brackets. Using this notation, we see that

$$P_g = \widetilde{g^2(t)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t)\, dt \tag{3.85a}$$

Note that the rms value of a signal is the square root of its mean square value. Therefore,

$$[g(t)]\mathrm{rms} = \sqrt{P_g} \tag{3.85b}$$

From Eqs. (3.82), it is clear that for a real signal $g(t)$, the time autocorrelation function $\mathcal{R}_g(\tau)$ is the time mean of $g(t)g(t+\tau)$. Thus,

$$\mathcal{R}_g(\tau) = \widetilde{g(t)g(t \pm \tau)} \tag{3.86}$$

This discussion also explains why we have been using the term time autocorrelation rather than just autocorrelation. This is to distinguish clearly the present autocorrelation function (a time average) from the statistical autocorrelation function (a statistical average) to be introduced in a future chapter.

Interpretation of Power Spectral Density

Because the PSD is a time average of the ESD of $g(t)$, we can argue along the lines used in the interpretation of ESD. We can readily show that the PSD $S_g(\omega)$ represents the power per unit

Table 3.3

$E_g = \int_{-\infty}^{\infty} g^2(t)\, dt$	$P_g = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t)\, dt = \lim_{T\to\infty} \frac{E_{g_T}}{T}$
$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t+\tau)\, dt$	$\mathcal{R}_g(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} g(t)g(t+\tau)\, dt = \lim_{T\to\infty} \frac{\psi_{g_T}(\tau)}{T}$
$\Psi_g(\omega) = \lvert G(\omega)\rvert^2$	$S_g(\omega) = \lim_{T\to\infty} \frac{\lvert G_T(\omega)\rvert^2}{T} = \lim_{T\to\infty} \frac{\Psi_{g_T}(\omega)}{T}$
$\psi_g(\tau) \Longleftrightarrow \Psi_g(\omega)$	$\mathcal{R}_g(\tau) \Longleftrightarrow S_g(\omega)$
$E_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_g(\omega)\, d\omega$	$P_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_g(\omega)\, d\omega$

bandwidth (in hertz) of the spectral components at the frequency ω. The power contributed by the spectral components within the band ω_1 to ω_2 is given by

$$\Delta P_g = \frac{1}{\pi} \int_{\omega_1}^{\omega_2} S_g(\omega)\, d\omega \tag{3.87}$$

Autocorrelation Method: A Powerful Tool

For a signal $g(t)$, the ESD, which is equal to $|G(\omega)|^2$, can also be found by taking the Fourier transform of its autocorrelation function. If the Fourier transform of a signal is enough to determine its ESD, then why do we needlessly complicate our lives by talking about autocorrelation functions? The reason for following this alternate route is to lay a foundation for dealing with power signals and random signals. The Fourier transform of a power signal generally does not exist. Moreover, the luxury of finding the Fourier transform is available only for deterministic signals, which can be described as functions of time. The random message signals that occur in communication problems (e.g., random binary pulse train) cannot be described as functions of time, and it is impossible to find their Fourier transforms. However, the autocorrelation function for such signals can be determined from their statistical information. This allows us to determine the PSD (the spectral information) of such a signal. Indeed, we may consider the autocorrelation approach as the generalization of Fourier techniques to power signals and random signals. The following example of a random binary pulse train dramatically illustrates the power of this technique.

EXAMPLE 3.23 Figure 3.42a shows a random binary pulse train $g(t)$. The pulse width is $T_b/2$, and one binary digit is transmitted every T_b seconds. A binary **1** is transmitted by the positive pulse, and a binary **0** is transmitted by the negative pulse. The two symbols are equally likely and occur randomly. We shall determine the autocorrelation function, the PSD, and the essential bandwidth of this signal.

We cannot describe this signal as a function of time because the precise waveform is not known due to its random nature. We do, however, know its behavior in terms of the averages (the statistical information). The autocorrelation function, being an average parameter (time average) of the signal, is determinable from the given statistical (average) information. We have [Eq. (3.82b)]

$$\mathcal{R}_g(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} g(t)g(t-\tau)\, dt$$

Figure 3.42b shows $g(t)$ by solid lines and $g(t-\tau)$, which is $g(t)$ delayed by τ, by dashed lines. To determine the integrand on the right-hand side of the above equation, we multiply $g(t)$ with $g(t-\tau)$, find the area under the product $g(t)g(t-\tau)$, and divide it by the averaging interval T. Let there be N bits (pulses) during this interval T so that $T = NT_b$, and as $T \to \infty$, $N \to \infty$. Thus,

$$\mathcal{R}_g(\tau) = \lim_{N\to\infty} \frac{1}{NT_b} \int_{-NT_b/2}^{NT_b/2} g(t)g(t-\tau)\, dt$$

Let us first consider the case of $\tau < T_b/2$. In this case there is an overlap (shown by the shaded region) between each pulse of $g(t)$ and that of $g(t-\tau)$. The area under

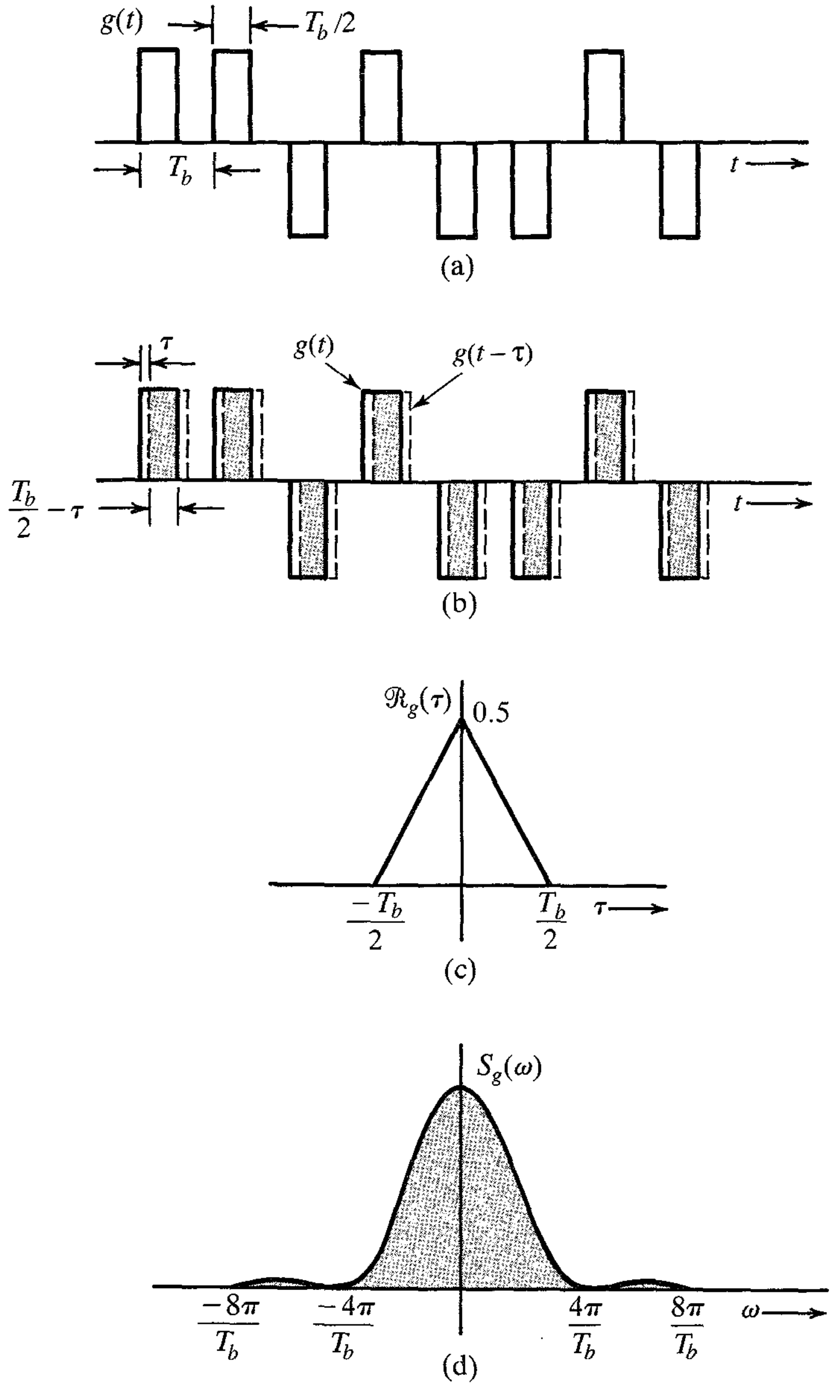

g(t)
T_b/2
T_b
t
(a)
τ
g(t)
g(t − τ)
T_b/2 − τ
t
(b)
ℛ_g(τ)
0.5
−T_b/2
T_b/2
τ
(c)
S_g(ω)
−8π/T_b
−4π/T_b
4π/T_b
8π/T_b
ω
(d)

Figure 3.42 Autocorrelation function and power spectral density function of a random binary pulse train.

the product $g(t)g(t-\tau)$ is $T_b/2-\tau$ for each pulse. Since there are N pulses during the averaging interval, then the total area under $g(t)g(t-\tau)$ is $N(T_b/2-\tau)$, and

$$\mathcal{R}_g(\tau) = \lim_{N\to\infty} \frac{1}{NT_b}\left[N\left(\frac{T_b}{2}-\tau\right)\right]$$

$$= \frac{1}{2}\left(1 - \frac{2\tau}{T_b}\right) \qquad \tau < \frac{T_b}{2}$$

Because $\mathcal{R}_g(\tau)$ is an even function of τ,

$$\mathcal{R}_g(\tau) = \frac{1}{2}\left(1 - \frac{2|\tau|}{T_b}\right) \qquad |\tau| < \frac{T_b}{2} \tag{3.88a}$$

as shown in Fig. 3.42c.

As we increase τ beyond $T_b/2$, there will be overlap between each pulse and its immediate neighbor. The two overlapping pulses are equally likely to be of the same polarity or of opposite polarity. Their product is equally likely to be 1 or -1 over the overlapping interval. On the average, half the pulse products will be 1 (positive-positive or negative-negative pulse combinations), and the remaining half pulse products will be -1 (positive-negative or negative-positive combinations). Consequently, the area under $g(t)g(t - \tau)$ will be zero when averaged over an infinitely large time ($T \to \infty$), and

$$\mathcal{R}_g(\tau) = 0 \qquad |\tau| > \frac{T_b}{2} \tag{3.88b}$$

The autocorrelation function in this case is the triangle function $\frac{1}{2}\Delta(t/T_b)$ shown in Fig. 3.42c. The PSD is the Fourier transform of $\frac{1}{2}\Delta(t/T_b)$, which is found in Example 3.15 (or Table 3.1, pair 19) as

$$S_g(\omega) = \frac{T_b}{4}\,\text{sinc}^2\left(\frac{\omega T_b}{4}\right) \tag{3.89}$$

The PSD is the square of the sinc function shown in Fig. 3.42d. From the result in Example 3.21, we conclude that the 90.28% of the area of this spectrum is contained within the band from 0 to $4\pi/T_b$ rad/s, or from 0 to $2/T_b$ Hz. Thus, the essential bandwidth may be taken as $2/T_b$ Hz (assuming a 90% power criterion). This example illustrates dramatically how the autocorrelation function can be used to obtain the spectral information of a (random) signal where conventional means of obtaining the Fourier spectrum are not usable.

Input and Output Power Spectral Densities

Because the PSD is a time average of ESDs, the relationship between the input and output signal PSDs of a linear time-invariant (LTI) system is similar to that of ESDs. Following the argument used for ESD [Eq. (3.74)], we can readily show that if $g(t)$ and $y(t)$ are the input and output signals of an LTI system with transfer function $H(\omega)$, then

$$S_y(\omega) = |H(\omega)|^2 S_g(\omega) \tag{3.90}$$

EXAMPLE 3.24 A noise signal $n_i(t)$ with PSD $S_{n_i}(\omega) = K$ is applied at the input of an ideal differentiator (Fig. 3.43a). Determine the PSD and the power of the output noise signal $n_o(t)$.

The transfer function of an ideal differentiator is $H(\omega) = j\omega$. If the noise at the demodulator output is $n_o(t)$, then from Eq. (3.90),

$$S_{n_o}(\omega) = |H(\omega)|^2\, S_{n_i}(\omega) = |j\omega|^2 K$$

The output PSD $S_{n_o}(\omega)$ is parabolic, as shown in Fig. 3.43c. The output noise power N_o is $1/2\pi$ times the area under the output PSD. Therefore,

$$N_o = \frac{1}{2\pi}\int_{-2\pi B}^{2\pi B} K\omega^2\, d\omega = K\int_{-2\pi B}^{2\pi B} \omega^2\, d\omega = \frac{8\pi^2 B^3 K}{3}$$

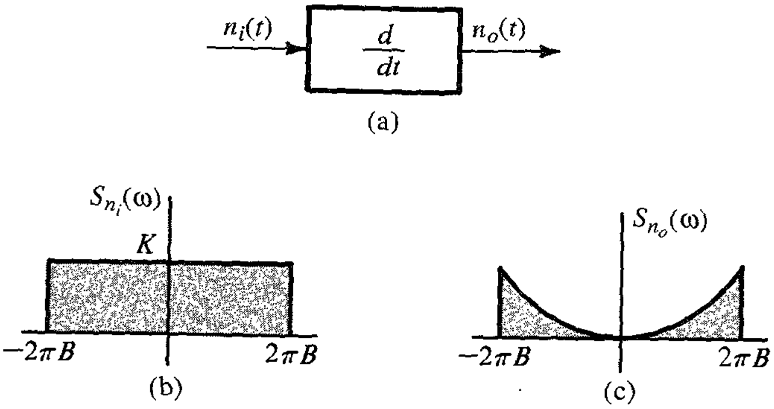


Figure 3.43 Power spectral densities at the input and the output of an ideal differentiator.

PSD of Modulated Signals

Following the argument in deriving Eqs. (3.69) and (3.70) for energy signals, we can derive similar results for power signals by taking the time averages. We can show that for a power signal $g(t)$, if

$$\varphi(t) = g(t)\cos\omega_0 t$$

then the PSD $S_\varphi(\omega)$ of the modulated signal $\varphi(t)$ is given by

$$S_\varphi(\omega) = \frac{1}{4}\left[S_g(\omega+\omega_0) + S_g(\omega-\omega_0)\right] \tag{3.91}$$

Thus, modulation shifts the PSD of $g(t)$ by $\pm\omega_0$. The power of $\varphi(t)$ is half the power of $g(t)$, that is,

$$P_\varphi = \frac{1}{2}P_g \qquad \omega_0 \geq 2\pi B \tag{3.92}$$

3.9 NUMERICAL COMPUTATION OF FOURIER TRANSFORM: THE DFT

To compute $G(\omega)$, the Fourier transform of $g(t)$, numerically, we have to use the samples of $g(t)$. Moreover, we can determine $G(\omega)$ only at some finite number of frequencies. Thus, we can only compute the samples of $G(\omega)$. For this reason, we shall now find the relationships between the samples of $g(t)$ and the samples of $G(\omega)$.

In numerical computations, the data must be finite. This means that the number of samples of $g(t)$ and $G(\omega)$ must be finite. In other words, we must deal with time-limited signals. If the signal is not time-limited, then we need to truncate it to make its duration finite. The same is true of $G(\omega)$. To begin with, let us consider a signal $g(t)$ of duration τ seconds, starting at $t = 0$, as shown in Fig. 3.44a. However, for reasons that will become clear as we go along, we shall consider the duration of $g(t)$ to be T_0, where $T_0 \geq \tau$, which makes $g(t) = 0$ in the interval $\tau < t \leq T_0$, as shown in Fig. 3.44a. Clearly, this makes no difference in the computation of $G(\omega)$. Let us take samples of $g(t)$ at uniform intervals of T_s seconds. There are a total of N_0 samples, where

$$N_0 = \frac{T_0}{T_s} \tag{3.93}$$

Now,*

$$\begin{aligned} G(\omega) &= \int_0^{T_0} g(t)e^{-j\omega t}\,dt \\ &= \lim_{T_s \to 0} \sum_{k=0}^{N_0-1} g(kT_s)e^{-j\omega kT_s}\,T_s \end{aligned} \tag{3.94}$$

Let us consider the samples of $G(\omega)$ at uniform intervals of ω_0. If G_r is the rth sample, that is, $G_r = G(r\omega_0)$, then from Eq. (3.94), we obtain

$$\begin{aligned} G_r &= \sum_{k=0}^{N_0-1} T_s g(kT_s)e^{-jr\omega_0 T_s k} \\ &= \sum_{k=0}^{N_0-1} g_k e^{-jr\Omega_0 k} \end{aligned} \tag{3.95}$$

where

$$g_k = T_s g(kT_s), \qquad G_r = G(r\omega_0), \qquad \Omega_0 = \omega_0 T_s \tag{3.96}$$

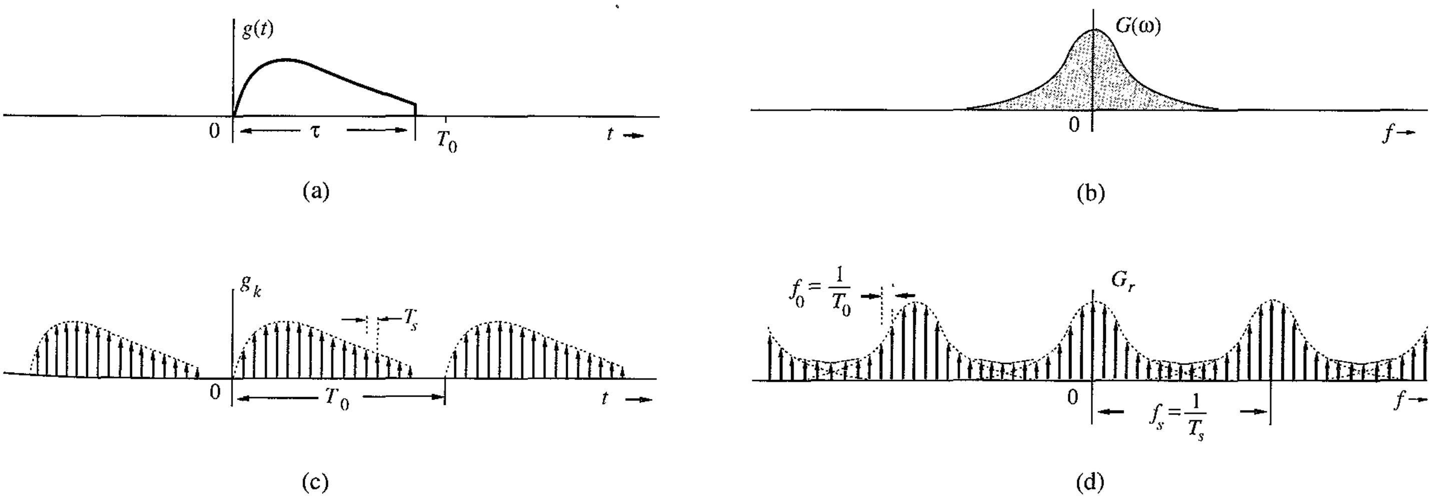


Figure 3.44 Relationship between the samples of $g(t)$ and of $G(\omega)$.

---

* The upper limit on the summation in Eq. (3.94) is $N_0 - 1$ (not N_0) because the last term in the sum starts at $(N_0 - 1)T_s$ and covers the area under the summand up to $N_0 T_s = T_0$.

Thus, Eq. (3.95) relates the samples of $g(t)$ to the samples of $G(\omega)$. In this derivation, we have assumed that $T_s \to 0$. In practice, it is not possible to make $T_s \to 0$ because it will increase the data enormously. We strive to make T_s as small as is practicable. This will result in some computational error.

We make an interesting observation from Eq. (3.95). The samples G_r are periodic with a period of $2\pi/\Omega_0$ samples. This follows from Eq. (3.95), which shows that $G_{(r+2\pi/\Omega_0)} = G_r$. Thus, only $2\pi/\Omega_0$ number of samples G_r can be independent. Equation (3.95) shows that G_r is determined by N_0 independent values g_k. Hence, for unique inverses of these equations, there can be only N_0 independent sample values G_r. This means

$$N_0 = \frac{2\pi}{\Omega_0} = \frac{2\pi}{\omega_0 T_s} = \frac{2\pi N_0}{\omega_0 T_0} \tag{3.97}$$

Hence,

$$\omega_0 = \frac{2\pi}{T_0} \quad \text{and} \quad f_0 = \frac{1}{T_0} \tag{3.98}$$

Thus, the spectral sampling interval ω_0 rad/sec. (or f_0 Hz) can be adjusted by a proper choice of T_0: the larger the T_0, the smaller the ω_0. The wisdom of selecting $T_0 \geq \tau$ is now clear. When T_0 is greater than τ, we shall have several zero-valued samples g_k in the interval from τ to T_0. Thus, by increasing the number of zero-valued samples of g_k, we reduce ω_0 [more closely spaced samples of $G(\omega)$], yielding more details of $G(\omega)$. This process of reducing ω_0 by the inclusion of zero-valued samples g_k is known as **zero padding**. Also, for a given sampling interval T_s, larger T_0 implies larger N_0. Thus, by selecting suitably large value of N_0, we can obtain samples of $G(\omega)$ as close as possible.

To find the inverse relationship, we multiply both sides of Eq. (3.95) by $e^{jm\Omega_0 r}$ and sum over r as

$$\sum_{r=0}^{N_0-1} G_r e^{jm\Omega_0 r} = \sum_{r=0}^{N_0-1} \left[\sum_{k=0}^{N_0-1} g_k e^{-jr\Omega_0 k} \right] e^{jm\Omega_0 r}$$

Interchanging the order of summation on the right-hand side,

$$\sum_{r=0}^{N_0-1} G_r e^{jm\Omega_0 r} = \sum_{k=0}^{N_0-1} g_k \left[\sum_{r=0}^{N_0-1} e^{j(m-k)\Omega_0 r} \right] \tag{3.99}$$

In order to find the inner sum on the right-hand side, we shall now show that

$$\sum_{k=0}^{N_0-1} e^{jn\Omega_0 k} = \begin{cases} N_0 & n = 0,\ \pm N_0,\ \pm 2N_0,\ \cdots \\ 0 & \text{otherwise} \end{cases} \tag{3.100}$$

To show this, recall that $\Omega_0 N_0 = 2\pi$ and $e^{jn\Omega_0 k} = 1$ for $n = 0, \pm N_0, \pm 2N_0, \cdots$, so that

$$\sum_{k=0}^{N_0-1} e^{jn\Omega_0 k} = \sum_{k=0}^{N_0-1} 1 = N_0 \qquad n = 0, \pm N_0, \pm 2N_0, \cdots$$

To compute the sum for other values of n, we note that the sum on the left-hand side of Eq. (3.100) is a geometric progression with common ratio $\alpha = e^{jn\Omega_0}$. Therefore (see Appendix E),

$$\sum_{k=0}^{N_0-1} e^{jn\Omega_0 k} = \frac{e^{jn\Omega_0 N_0} - 1}{e^{jn\Omega_0} - 1} = 0 \qquad e^{jn\Omega_0 N_0} = e^{j2\pi n} = 1$$

This proves Eq. (3.100). It now follows that the inner sum on the right-hand side of Eq. (3.99) is zero for $k \neq m$, and the sum is N_0 when $k = m$. Therefore, the outer sum will have only one nonzero term when $k = m$, and it is $N_0 g_k = N_0 g_m$. Therefore,

$$g_m = \frac{1}{N_0} \sum_{r=0}^{N_0-1} G_r e^{jm\Omega_0 r} \qquad \Omega_0 = \frac{2\pi}{N_0} \tag{3.101}$$

Equation (3.101) reveals the interesting fact that $g_{(m+N_0)} = g_m$. This means that the sequence g_k is also periodic with a period of N_0 samples (representing the time duration $N_0 T_s = T_0$ seconds). Moreover, G_r is also periodic with a period of N_0 samples, representing a frequency interval $N_0 \omega_0 = (T_0/T_s)(2\pi/T_0) = 2\pi/T_s = \omega_s$ rad/s. This is equal to $1/T_s$ Hz. But $1/T_s$ is the number of samples of $g(t)$ per second. Thus, $1/T_s = f_s$ is the sampling frequency (in hertz) of $g(t)$. This means G_r is N_0-periodic, repeating every f_s Hz. Let us summarize the results derived so far. We have proved the discrete Fourier transform (DFT) pair

$$G_r = \sum_{k=0}^{N_0-1} g_k e^{-jr\Omega_0 k} \tag{3.102a}$$

$$g_k = \frac{1}{N_0} \sum_{r=0}^{N_0-1} G_r e^{jk\Omega_0 r} \tag{3.102b}$$

where

$$\begin{aligned} g_k &= T_s g(kT_s) & G_r &= G(r\omega_0) \\ \omega_0 &= \frac{2\pi}{T_0} = 2\pi f_0 & \omega_s &= \frac{2\pi}{T_s} = 2\pi f_s \\ N_0 &= \frac{T_0}{T_s} = \frac{\omega_s}{\omega_0} = \frac{f_s}{f_0} & \Omega_0 &= \omega_0 T_s = \frac{2\pi}{N_0} \end{aligned} \tag{3.103}$$

Both the sequences g_k and G_r are periodic with a period of N_0 samples. This results in g_k repeating with period T_0 seconds and G_r repeating with period $\omega_s = 2\pi/T_s$ rad/s, or $f_s = 1/T_s$ Hz (the sampling frequency). The sampling interval of g_k is T_s seconds and the sampling interval of G_r is $\omega_0 = 2\pi/T_0$ rad/s, or $f_0 = 1/T_0$ Hz. This is shown in Fig. 3.44c and d. For convenience, we have used the frequency variable f (in hertz) rather than ω (in radious per second).

We have assumed $g(t)$ to be time-limited to τ seconds. This makes $G(\omega)$ non-band-limited.* Hence, the periodic repetition of the spectra G_r, as shown in Fig. 3.44d, will cause overlapping of spectral components, resulting in error. The nature of this error, known as **aliasing error**, is explained in more detail in Chapter 6. The spectrum G_r repeats every f_s Hz. The aliasing error is reduced by increasing f_s, the repetition frequency (see Fig. 3.44d). To summarize, the computation of G_r using DFT has aliasing error when $g(t)$ is time-limited. This error can be made as small as desired by increasing the sampling frequency $f_s = 1/T_s$ (or reducing the sampling interval T_s). The aliasing error is the direct result of the nonfulfillment of the requirement $T_s \to 0$ in Eq. (3.94).

---

* We can show that a signal cannot be simultaneously time-limited and band-limited. If it is one, it cannot be the other, and vice versa.[3]

When $g(t)$ is not time-limited, we need to truncate it to make it time-limited. This will cause further error in G_r. This error can be reduced as much as desired by appropriately increasing the truncating interval T_0.*

In computation of the inverse Fourier transform [by using the inverse DFT in Eq. (3.102b)] we have similar problems. If $G(\omega)$ is band-limited, $g(t)$ is not time-limited, and the periodic repetition of samples g_k will overlap (aliasing in the time domain). We can reduce the aliasing error by increasing T_0, the period of g_k (in seconds). This is equivalent to reducing the frequency sampling interval $f_0 = 1/T_0$ of $G(\omega)$. Moreover, if $G(\omega)$ is not band-limited, we need to truncate it. This will cause an additional error in the computation of g_k. By increasing the truncation bandwidth, we can reduce this error. In practice, (tapered) window functions are often used for truncation[5] in order to reduce the severity of some problems caused by straight truncation (also known as rectangular windowing).

Because G_r is N_0-periodic, we need to determine the values of G_r over any one period. It is customary to determine G_r over the range $(0,\ N_0 - 1)$ rather than over the range $(-N_0/2,\ N_0/2 - 1)$. The identical remark applies to g_k.

Choice of T_s, T_0, and N_0

In DFT computation, we first need to select suitable values for N_0, T_s, and T_0. For this purpose we should first decide on B, the essential bandwidth of $g(t)$. From Fig. 3.44d, it is clear that the spectral overlapping (aliasing) occurs at the frequency $f_s/2$ Hz. This spectral overlapping may also be viewed as the spectrum beyond $f_s/2$ folding back at $f_s/2$. Hence, this frequency is also called the **folding frequency**. If the folding frequency is chosen such that the spectrum $G(\omega)$ is negligible beyond the folding frequency, aliasing (the spectral overlapping) is not significant. Hence, the folding frequency should at least be equal to the highest significant frequency, that is , the frequency beyond which $G(\omega)$ is negligible. We shall call this frequency the **essential bandwidth** B (in hertz). If $g(t)$ is band-limited, then clearly, its bandwidth is identical to the essential bandwidth. Thus,

$$\frac{f_s}{2} \geq B \tag{3.104a}$$

Moreover, the sampling interval $T_s = 1/f_s$ [Eq. (3.103)]. Hence,

$$T_s \leq \frac{1}{2B} \tag{3.104b}$$

Once we pick B, we can choose T_s according to Eq. (3.104b). Also,

$$f_0 = \frac{1}{T_0} \tag{3.105}$$

where f_0 is the **frequency resolution** [separation between samples of $G(\omega)$]. Hence, if f_0 is given, we can pick T_0 according to Eq. (3.105). Knowing T_0 and T_s, we determine N_0 from

$$N_0 = \frac{T_0}{T_s} \tag{3.106}$$

In general, if the signal is time-limited, $G(\omega)$ is not band-limited, and there is aliasing in the computation of G_r. To reduce the aliasing effect, we need to increase the folding frequency, that is, reduce T_s (the sampling interval) as much as is practicable. If the signal is band-limited, $g(t)$ is not time-limited, and there is aliasing (overlapping) in the computation of g_k. To reduce this aliasing, we need to increase T_0, the period of g_k. This results in reducing the frequency sampling interval f_0 (in hertz). In either case (reducing T_s in the time-limited case or increasing T_0 in the band-limited case), for higher accuracy, we need to increase the number of samples N_0 because $N_0 = T_0/T_s$. There are also signals that are neither time-limited nor band-limited. In such cases, we need to reduce T_s and increase T_0.

---

* The DFT relationships represent a transform in their own right, and they are exact. If, however, we identify g_k and G_r as the samples of a signal $g(t)$ and its Fourier transform $G(\omega)$, respectively, then the DFT relationships are approximations because of the aliasing and truncating effects.

Points of Discontinuity

If $g(t)$ has a jump discontinuity at a sampling point, the sample value should be taken as the average of the values on the two sides of the discontinuity because the Fourier representation at a point of discontinuity converges to the average value.

DFT Computations Using the FFT Algorithm

The number of computations required in performing the DFT was dramatically reduced by an algorithm developed by Tukey and Cooley in 1965.[6] This algorithm, known as the **fast Fourier transform (FFT)**, reduces the number of computations from something on the order of N_0^2 to $N_0 \log N_0$. To compute one sample G_r from Eq. (3.102a), we require N_0 complex multiplications and $N_0 - 1$ complex additions. To compute N_0 values of G_r ($r = 0, 1, \cdots, N_0 - 1$), we require a total of N_0^2 complex multiplications and $N_0(N_0 - 1)$ complex additions. For large N_0, this can be prohibitively time-consuming, even for a very high-speed computer. The FFT is, thus, a life saver in signal processing applications. The FFT algorithm is simplified if we choose N_0 to be a power of 2, although this is not necessary, in general. Details of the FFT can be found in any book on signal processing.[3]

Let us consider two examples illustrating the use of DFT in finding the Fourier transform. We shall use MATLAB to find DFT by the FFT algorithm. In the first example, the signal $g(t) = e^{-2t}u(t)$ starts at $t = 0$. In the second example, we use $g(t) = \text{rect}(t)$, which starts at $t = -\frac{1}{2}$.

Computer Example C3.1

Use DFT (implemented by the FFT algorithm) to compute the Fourier transform of $e^{-2t}u(t)$. Plot the resulting Fourier spectra.

We first determine T_s and T_0. The Fourier transform of $e^{-2t}u(t)$ is $1/(j\omega+2)$. This low-pass signal is not band-limited. Let us take its essential bandwidth to be that frequency where $|G(\omega)|$ becomes 1% of its peak value, which occurs at $\omega = 0$. Observe that

$$|G(\omega)| = \frac{1}{\sqrt{\omega^2+4}} \approx \frac{1}{\omega} \qquad \omega \gg 2$$

Also, the peak of $|G(\omega)|$ is at $\omega = 0$, where $|G(0)| = 0.5$. Hence, the essential bandwidth B is at $\omega = 2\pi B$, where

$$|G(\omega)| \approx \frac{1}{2\pi B} = 0.5 \times 0.01 \quad \Rightarrow B = \frac{100}{\pi}\text{ Hz}$$

and from Eq. (3.104b),

$$T_s \le \frac{1}{2B} = 0.005\pi = 0.0157$$

Let us round this value down to $T_s = 0.015625$ second so that we have 64 samples per second. The second issue is to determine T_0. The signal is not time-limited. We need to truncate it at T_0 such that $g(T_0) \ll 1$. We shall pick $T_0 = 4$ (eight time constants of the signal), which yields $N_0 = T_0/T_s = 256$. This is a power of 2. Note that there is a great deal of flexibility in determining T_s and T_0, depending on the accuracy desired and the computational capacity available. We could just as well have picked $T_0 = 8$ and $T_s = 1/32$, yielding $N_0 = 256$, although this would have given a slightly higher aliasing error.

Because the signal has a jump discontinuity at $t = 0$, the first sample (at $t = 0$) is 0.5, the averages of the values on the two sides of the discontinuity. The MATLAB program, which implements the DFT using the FFT algorithm is as follows:

```
Ts=1/64; T0=4; N0=T0/Ts;
t=0:Ts:Ts*(N0-1);t=t';
g=Ts*exp(-2*t);
g(1)=Ts*0.5;
G=fft(g);
[Gp,Gm]=cart2pol(real(G),imag(G));
k=0:N0-1; k=k';
w=2*pi*k/T0;
subplot(211),stem(w(1:32),Gm(1:32));
subplot(212),stem(w(1:32),Gp(1:32))
```

Because G_r is N_0-periodic, $G_r = G_{(r+256)}$ so that $G_{256} = G_0$. Hence, we need to plot G_r over the range $r = 0$ to 255 (not 256). Moreover, because of this periodicity, $G_{-r} = G_{(-r+256)}$,

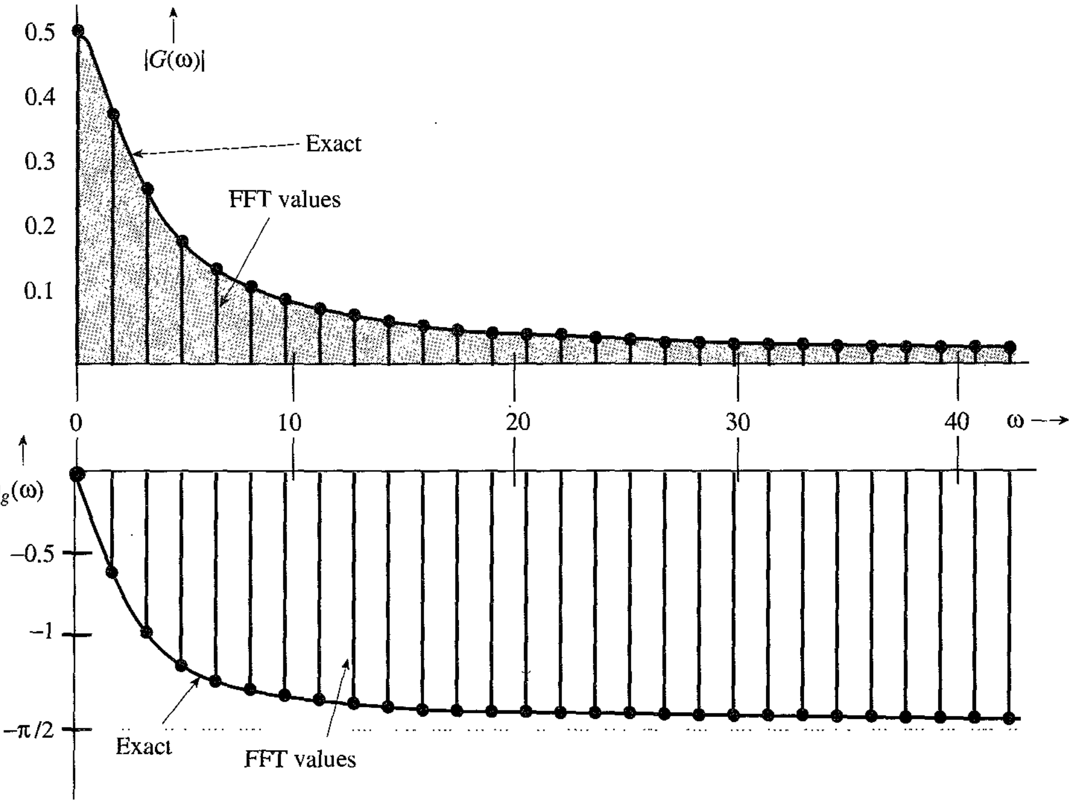


Figure 3.45 Discrete Fourier transform of an exponential signal $e^{-2t}u(t)$.

and the G_r over the range of $r = -127$ to -1 are identical to the G_r over the range of $r = 129$ to 255. Thus, $G_{-127} = G_{129}$, $G_{-126} = G_{130}$, $\cdots$, $G_{-1} = G_{255}$. In addition, because of the property of conjugate symmetry of the Fourier transform, $G_{-r} = G_r^*$, it follows that $G_{129} = G_{127}^*$, $G_{130} = G_{126}^*$, $\cdots$, $G_{255} = G_1^*$. Thus, the plots beyond $r = N_0/2$ (128 in this case) are not necessary for real signals (because they are conjugates of G_r for $r = 0$ to 128.

The plot of the Fourier spectra in Fig. 3.45 shows the samples of magnitude and phase of $G(\omega)$ at the intervals of $1/T_0 = 1/4$ Hz or $\omega_0 = 1.5708$ rad/s. In Fig. 3.45, we have shown only the first 28 points (rather than all 128 points) to avoid too much crowding of the data.

In this example, we knew $G(\omega)$ beforehand and hence could make intelligent choices for B (or the sampling frequency f_s). In practice, we generally do not know $G(\omega)$ beforehand. In fact, that is the very thing we are trying to determine. In such a case, we must make an intelligent guess for B or f_s from circumstantial evidence. We should then continue reducing the value of T_s and recomputing the transform until the result stabilizes within the desired number of significant digits.

Next, we compute the Fourier transform of $g(t) = 8$ rect (t).

Computer Example C3.2

Use DFT (implemented by the FFT algorithm) to compute the Fourier transform of 8 rect(t). Plot the resulting Fourier spectra.

This gate function and its Fourier transform are shown in Fig. 3.46a and b. To determine the value of the sampling interval T_s, we must first decide on the essential bandwidth B. From Fig. 3.46b, we see that $G(\omega)$ decays rather slowly with ω. Hence, the essential bandwidth B is rather large. For instance, at $B = 15.5$ Hz (97.39 rad/s), $G(\omega) = -0.1643$, which is about 2% of the peak at $G(0)$. Hence, the essential bandwidth may be taken as 16 Hz. However, we shall deliberately take $B = 4$ for two reasons: (1) to show the effect of aliasing and (2) the use of $B > 4$ will give an enormous number of samples, which cannot be conveniently displayed on the book-sized sheet without losing sight of the essentials. Thus, we shall intentionally accept approximation in order to clarify the concepts of DFT graphically.

The choice of $B = 4$ results in the sampling interval $T_s = 1/2B = \frac{1}{8}$. Looking again at the spectrum in Fig. 3.46b, we see that the choice of the frequency resolution $f_0 = \frac{1}{4}$ Hz is reasonable. This will give four samples in each lobe of $G(\omega)$. In this case $T_0 = 1/f_0 = 4$ seconds and $N_0 = T_0/T = 32$. The duration of $g(t)$ is only 1 second. We must repeat it every 4 seconds ($T_0 = 4$), as shown in Fig. 3.46c, and take samples every $\frac{1}{8}$ second. This gives us 32 samples ($N_0 = 32$). Also,

$$\begin{aligned} g_k &= T_s g(kT) \\ &= \frac{1}{8} g(kT) \end{aligned}$$

Since $g(t) = 8$ rect (t), the values of g_k are 1, 0, or 0.5 (at the points of discontinuity), as shown in Fig. 3.46c. In this figure, g_k is shown as a function of t as well as k, for convenience.

In the derivation of the DFT, we assumed that $g(t)$ begins at $t = 0$ (Fig. 3.44a), and then took N_0 samples over the interval $(0, T_0)$. In the present case, however, $g(t)$ begins at $-\frac{1}{2}$. This difficulty is easily resolved when we realize that the DFT found by this procedure is actually the DFT of g_k repeating periodically every T_0 seconds. From Fig. 3.46c, it is clear that repeating the segment of g_k over the interval from -2 to 2 seconds periodically is identical to repeating the segment of g_k over the interval from 0 to 4 seconds. Hence, the DFT of the samples taken from -2 to 2 seconds is the same as that of the samples taken from 0 to 4 seconds. Therefore, regardless of where $g(t)$

starts, we can always take the samples of $g(t)$ and its periodic extension over the interval from 0 to T_0. In the present example, the 32 sample values are

$$g_k = \begin{cases} 1 & 0 \le k \le 3 \quad \text{and} \quad 29 \le k \le 31 \\ 0 & 5 \le k \le 27 \\ 0.5 & k = 4, 28 \end{cases}$$

Observe that the last sample is at $t = 31/8$, not at 4, because the signal repetition starts at $t = 4$, and the sample at $t = 4$ is the same as the sample at $t = 0$. Now, $N_0 = 32$ and $\Omega_0 = 2\pi/32 = \pi/16$. Therefore [see Eq. (3.102a)],

$$G_r = \sum_{k=0}^{31} g_k e^{-jr\frac{\pi}{16}k}$$

The MATLAB program, which implements this DFT equation using the FFT algorithm, is given next. First we write a MATLAB program to generate 32 samples of g_k, and then we compute the DFT.

```
% (c32.m)
B=4; f0=1/4;
Ts=1/(2*B); T0=1/f0;
N0=T0/Ts;
k=0:N0; k=k';
for m=1:length(k)
   if k(m)>=0 & k(m)<=3, gk(m)=1; end
   if k(m)==4 & k(m)==28 gk(m)=0.5; end
   if k(m)>=5 & k(m)<=27, gk(m)=0; end
   if k(m)>=29 & k(m)<=31, gk(m)=1; end
end
gk=gk';
Gr=fft(gk);
subplot(211),stem(k,gk)
subplot(212),stem(k,Gr)
```

Figure 3.46d shows the plot of G_r.

The samples G_r are separated by $f_0 = 1/T_0$ Hz. In this case $T_0 = 4$, so the frequency resolution f_0 is $\frac{1}{4}$ Hz, as desired. The folding frequency $f_s/2 = B = 4$ Hz corresponds to $r = N_0/2 = 16$. Because G_r is N_0-periodic ($N_0 = 32$), the values of G_r for $r = -16$ to $n = -1$ are the same as those for $r = 16$ to $n = 31$. The DFT gives us the samples of the spectrum $G(\omega)$.

For the sake of comparison, Fig. 3.46d also shows the shaded curve 8 sinc$(\omega/2)$, which is the Fourier transform of 8 rect(t). The values of G_r computed from the DFT equation show aliasing error, which is clearly seen by comparing the two superimposed plots. The error in G_2 is just about 1.3%. However, the aliasing error increases rapidly with r. For instance, the error in G_6 is about 12%, and the error in G_{10} is 33%. The error in G_{14} is a whopping 72%. The percent error increases rapidly near the folding frequency ($r = 16$) because $g(t)$ has a jump discontinuity, which makes $G(\omega)$ decay slowly as $1/\omega$. Hence, near the folding frequency, the inverted tail (due to aliasing) is very nearly equal to $G(\omega)$ itself. Moreover, the final values are the difference between the exact and the folded values (which are very close to the exact values). Hence, the percent error near the folding frequency ($r = 16$ in this case) is very high, although the absolute error is very small. Clearly, for signals with jump discontinuities, the aliasing error near the folding frequency

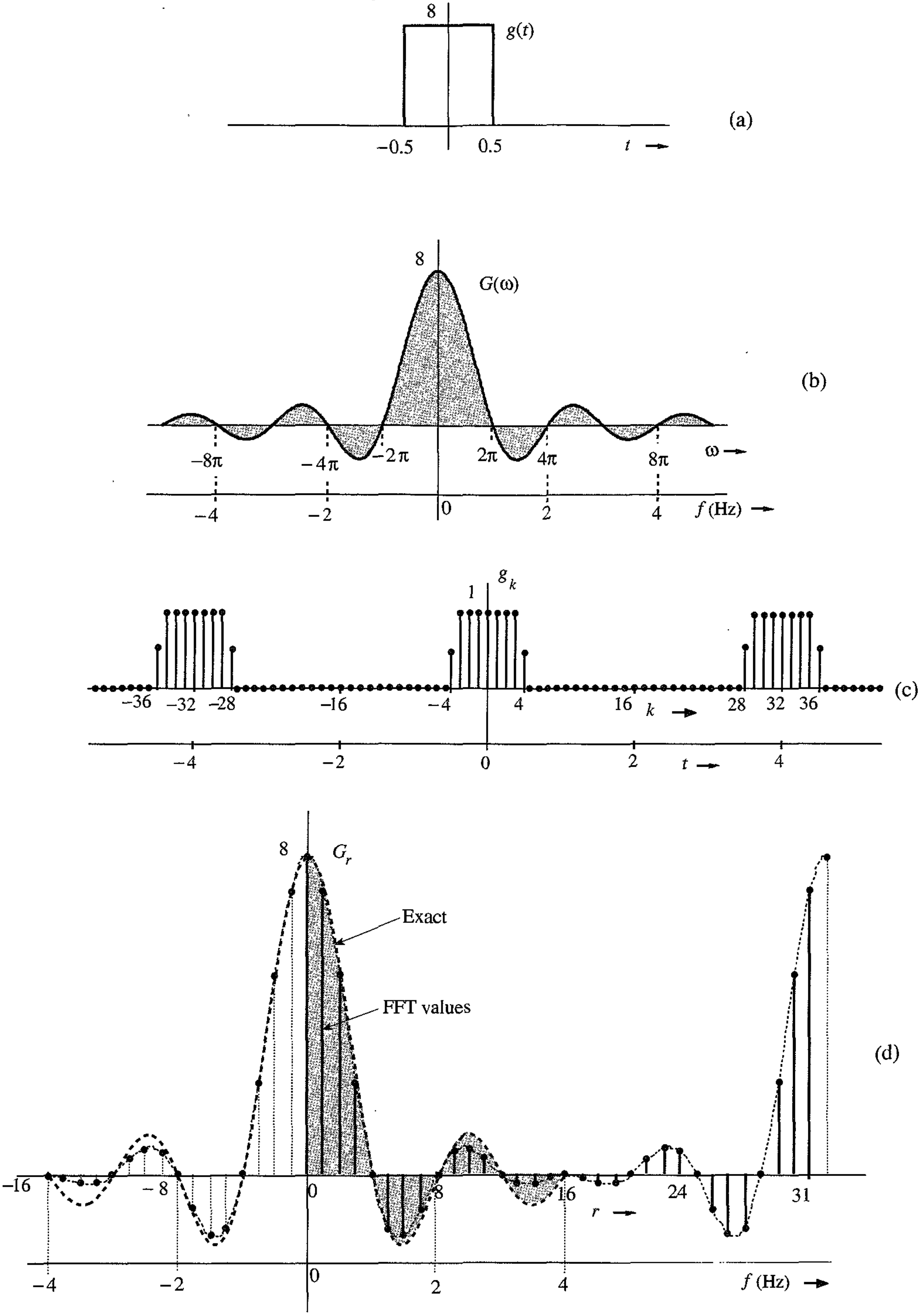


Figure 3.46 Discrete Fourier transform of a gate pulse.

will always be high (in percentage terms), regardless of the choice of N_0. To ensure a negligible aliasing error at any value r, we must make sure that $N_0 \gg r$. This observation is valid for all signals with jump discontinuities.

Filtering

We generally think of filtering in terms of some hardware-oriented solution (namely, building a circuit with RLC components and operational amplifiers). However, filtering also has a software-oriented solution [a computer algorithm that yields the filtered output $y(t)$ for a given input $g(t)$]. This can be conveniently accomplished by using the DFT. If $g(t)$ is the signal to be filtered, then G_r, the DFT of g_k, is found. The spectrum G_r is then shaped (filtered) as desired by multiplying G_r by H_r, where H_r are the samples of the filter transfer function $H(\omega)$ $[H_r = H(r\omega_0)]$. Finally, we take the IDFT of $G_r H_r$ to obtain the filtered output y_k $[y_k = T_s y(kT)]$. This procedure is demonstrated in the following example.

Computer Example C3.3

The signal $g(t)$ in Fig. 3.47a is passed through an ideal low-pass filter of transfer function $H(\omega)$, shown in Fig. 3.47b. Using DFT, find the filter output.

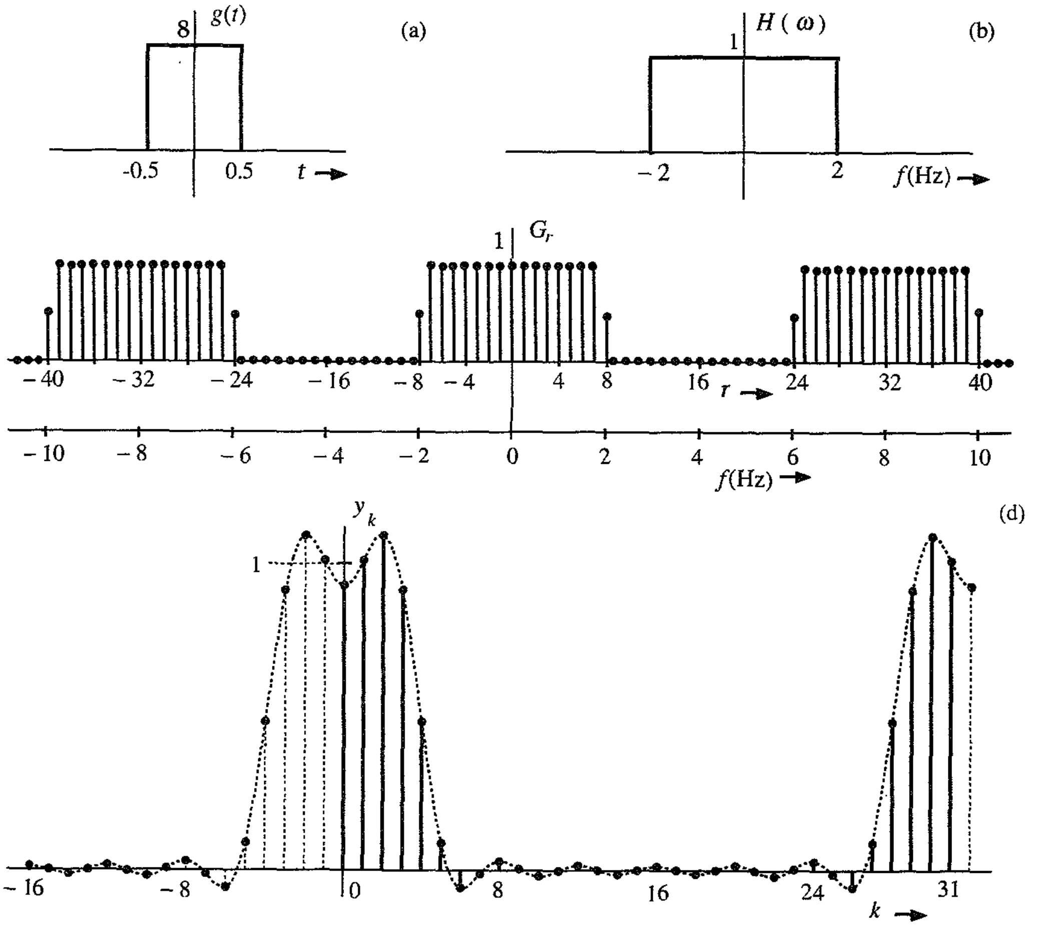


Figure 3.47 Filtering $g(t)$ through $H(\omega)$.

We have already found the 32-point DFT of $g(t)$ (see Fig. 3.46d). Next we multiply G_r by H_r. To compute H_r, we remember that in computing the 32-point DFT of $g(t)$, we have used $f_0 = \frac{1}{4}$. Because G_r is 32-periodic, H_r must also be 32-periodic with samples separated by $\frac{1}{4}$ Hz. This means that H_r must be repeated every 8 Hz or 16π rad/s (see Fig. 3.47c). This gives the 32 samples of H_r over $0 \le \omega \le 16\pi$ as follows:

$$H_r = \begin{cases} 1 & 0 \le r \le 7 \quad \text{and} \quad 25 \le r \le 31 \\ 0 & 9 \le r \le 23 \\ 0.5 & r = 8, 24 \end{cases}$$

We multiply G_r by H_r and take the inverse DFT. The resulting output signal is shown in Fig. 3.47d. Table 3.4 gives a printout of g_k, G_r, H_r, Y_r, and y_k.

We have already found the 32-point DFT (G_r) of $g(t)$ in Example C3.2. The MATLAB program of Example C3.2 should be saved as an m-file, e.g., "c32.m." We can import G_r in the MATLAB environment by the command "c32." Next, we generate 32-point samples of H_r, multiply G_r by H_r, and take the inverse DFT to compute y_k. We can also find y_k by convolving g_k with h_k.

```
c32;
r=0:32; r=r';
for m=1:length(r)
   if r(m)>=0 & r(m)<=7, Hr(m)=1; end
   if r(m)>=25 & r(m)<=31, Hr(m)=1; end
   if r(m)>=9 & r(m)<=23, Hr(m)=0; end
   if r(m)==8 & r(m)==24, Hr(m)=0.5; end
```

Table 3.4

No.	g_k	G_r	H_r	G_rH_r	y_k
0	1	8.000	1	8.000	.9285
1	1	7.179	1	7.179	1.009
2	1	5.027	1	5.027	1.090
3	1	2.331	1	2.331	.9123
4	0.5	0.000	1	0.000	.4847
5	0	−1.323	1	−1.323	.08884
6	0	−1.497	1	−1.497	−.05698
7	0	−.8616	1	−.8616	−.01383
8	0	0.000	0.5	0.000	.02933
9	0	.5803	0	0.000	.004837
10	0	.6682	0	0.000	−.01966
11	0	.3778	0	0.000	−.002156
12	0	0.000	0	0.000	.01534
13	0	−.2145	0	0.000	.0009828
14	0	−.1989	0	0.000	−.01338
15	0	−.06964	0	0.000	−.0002876
16	0	0.000	0	0.000	.01280
17	0	−.06964	0	0.000	.0002876
18	0	−.1989	0	0.000	−.01338
19	0	−.2145	0	0.000	.0009828
20	0	0.000	0	0.000	.01534
21	0	.3778	0	0.000	−.002156
22	0	.6682	0	0.000	−.01966
23	0	.5803	0	0.000	.004837
24	0	0.000	0.5	0.000	.03933
25	0	−.8616	1	−.8616	−.01383
26	0	−1.497	1	−1.497	−.05698
27	0	−1.323	1	−1.323	.08884
28	0.5	0.000	1	0.000	.4847
29	1	2.331	1	2.331	.9123
30	1	5.027	1	5.027	1.090
31	1	7.179	1	7.179	1.009

```
end
Hr=Hr';
Yr=Gr.*Hr;
yk=ifft(Yr);
clg,stem(k,yk)
```

REFERENCES

1. R. V. Churchill, and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed., McGraw-Hill, New York, 1978.
2. R. N. Bracewell, *Fourier Transform and Its Applications*, rev. 2nd ed., McGraw-Hill, New York, 1986.
3. B. P. Lathi, *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, Carmichael, CA, 1998.
4. E. A. Guillemin, *Theory of Linear Physical Systems*, Wiley, New York, 1963.
5. F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE*, vol. 66, pp. 51–83, January 1978.
6. Tukey and Cooley, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation*, Vol. 19, pp. 297–301, April, 1965.

PROBLEMS

3.1-1 Show that the Fourier transform of $g(t)$ may be expressed as

$$G(\omega) = \int_{-\infty}^{\infty} g(t)\cos\omega t\, dt - j\int_{-\infty}^{\infty} g(t)\sin\omega t\, dt$$

Hence, show that if $g(t)$ is an even function of t, then

$$G(\omega) = 2\int_{0}^{\infty} g(t)\cos\omega t\, dt$$

and if $g(t)$ is an odd function of t, then

$$G(\omega) = -2j\int_{0}^{\infty} g(t)\sin\omega t\, dt$$

Hence, prove that:

If $g(t)$ is:	*Then $G(\omega)$ is*:
a real and even function of t	a real and even function of ω
a real and odd function of t	an imaginary and odd function of ω
an imaginary and even function of t	an imaginary and even function of ω
a complex and even function of t	a complex and even function of ω
a complex and odd function of t	a complex and odd function of ω

3.1-2 **(a)** Show that for a real $g(t)$, the inverse transform, Eq. (3.8b), can be expressed as

$$g(t) = \frac{1}{\pi}\int_{0}^{\infty} |G(\omega)|\cos[\omega t + \theta_g(\omega)]\, d\omega$$

This is the trigonometric form of the (inverse) Fourier transform. Compare this with the compact trigonometric Fourier series.

(b) Express the Fourier integral (inverse Fourier transform) for $g(t) = e^{-at}u(t)$ in the trigonometric form given in part **(a)**.

3.1-3 If $g(t) \Longleftrightarrow G(\omega)$, then show that $g^*(t) \Longleftrightarrow G^*(-\omega)$.

3.1-4 From definition (3.8a), find the Fourier transforms of the signals $g(t)$ shown in Fig. P3.1-4.

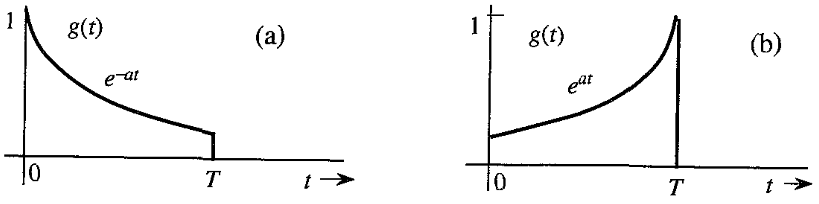

1
g(t)
e^{-at}
(a)
0
T
t →
1
g(t)
e^{at}
(b)
0
T
t →

Figure P3.1-4

3.1-5 From definition (3.8a), find the Fourier transforms of the signals shown in Fig. P3.1-5.

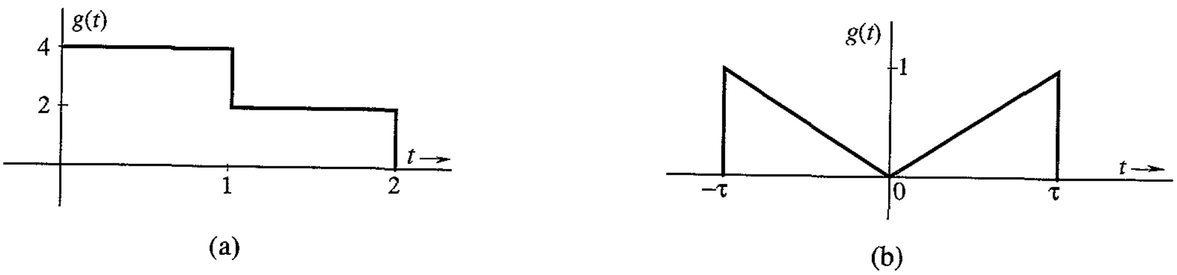

g(t)
4
2
1
2
t →
(a)
g(t)
1
−τ
0
τ
t →
(b)

Figure P3.1-5

3.1-6 From definition (3.8b), find the inverse Fourier transforms of the spectra shown in Fig. P3.1-6.

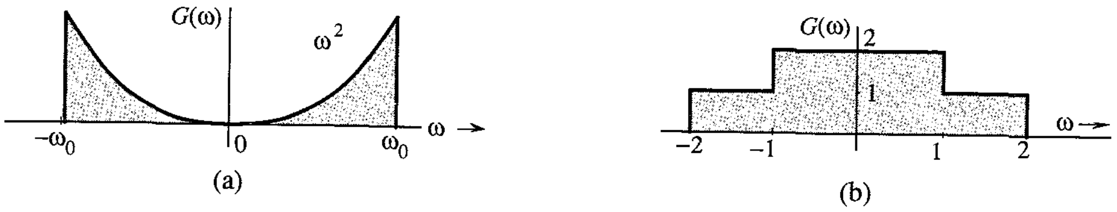

G(ω)
ω²
−ω₀
0
ω₀
ω →
(a)
G(ω)
2
1
−2
−1
1
2
ω →
(b)

Figure P3.1-6

3.1-7 From definition (3.8b), find the inverse Fourier transforms of the spectra shown in Fig. P3.1-7.

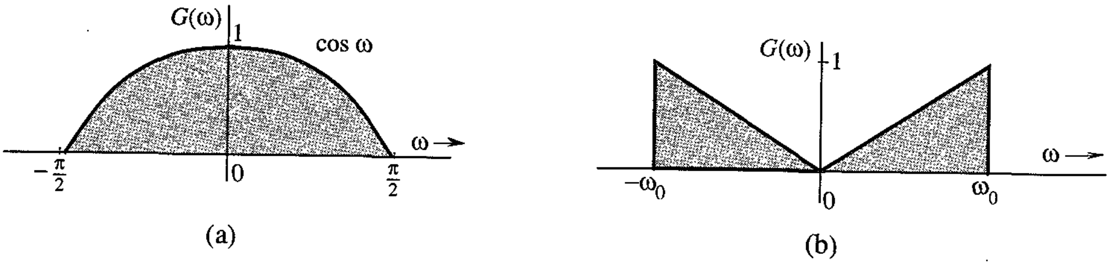
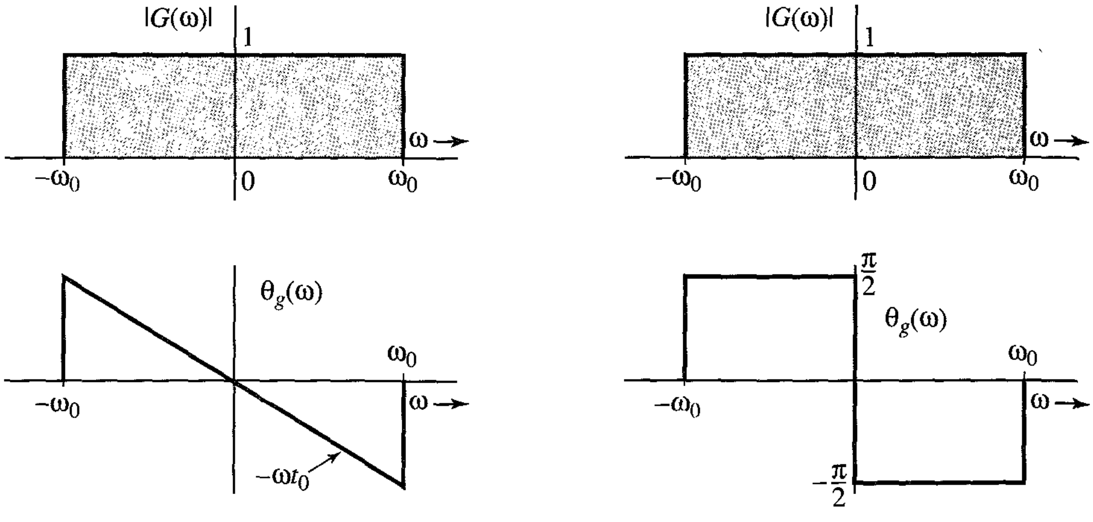

G(ω)
1
cos ω
−π/2
0
π/2
ω →
(a)
G(ω)
1
−ω₀
0
ω₀
ω →
(b)

Figure P3.1-7

3.1-8 Find the inverse Fourier transform of $G(\omega)$ for the spectra shown in Fig. P3.1-8.



(a) (b)

Figure P3.1-8

Hint: $G(\omega) = |G(\omega)|e^{j\theta_g(\omega)}$. For part **(a)**, $G(\omega) = 1e^{-j\omega t_0}$, $|\omega| \le \omega_0$. For part **(b)**,

$$G(\omega) = \begin{cases} 1e^{-j\pi/2} = -j & 0 < \omega \le \omega_0 \\ 1e^{j\pi/2} = j & 0 > \omega \ge -\omega_0 \end{cases}$$

This problem illustrates how different phase spectra (both with the same amplitude spectrum) represent entirely different signals.

3.2-1 Sketch the following functions:

(a) rect $(t/2)$; **(b)** $\Delta(3\omega/100)$; **(c)** rect $(t - 10/8)$; **(d)** sinc $(\pi\omega/5)$; **(e)** sinc $[(\omega - 10\pi)/5]$; **(f)** sinc $(t/5)$ rect $(t/10\pi)$. *Hint*: $g(\frac{x-a}{b})$ is $g(\frac{x}{b})$ right-shifted by a.

3.2-2 From definition (3.8a), show that the Fourier transform of rect $(t - 5)$ is sinc $(\omega/2)e^{-j5\omega}$.

3.2-3 From definition (3.8b), show that the inverse Fourier transform of rect $[(\omega - 10)/2\pi]$ is sinc $(\pi t)\, e^{j10t}$.

3.2-4 Using pairs 7 and 12 (Table 3.1) show that $u(t) \Longleftrightarrow \pi\delta(\omega) + 1/j\omega$.

3.2-5 Show that cos $(\omega_0 t + \theta) \Longleftrightarrow \pi[\delta(\omega + \omega_0)e^{-j\theta} + \delta(\omega - \omega_0)e^{j\theta}]$. *Hint*: Express cos $(\omega_0 t + \theta)$ in terms of exponentials using Euler's formula.

3.3-1 Apply the symmetry property to the appropriate pair in Table 3.1 to show that:

(a) $0.5[\delta(t) + (j/\pi t)]$

(b) $\delta(t + T) + \delta(t - T) \Longleftrightarrow 2\cos T\omega$;

(c) $\delta(t + T) - \delta(t - T) \Longleftrightarrow 2j\sin T\omega$. *Hint*: $g(-t) \Longleftrightarrow G(-\omega)$ and $\delta(t) = \delta(-t)$.

3.3-2 The Fourier transform of the triangular pulse $g(t)$ in Fig. P3.3-2a is given as

$$G(\omega) = \frac{1}{\omega^2}(e^{j\omega} - j\omega e^{j\omega} - 1)$$

Using this information, and the time-shifting and time-scaling properties, find the Fourier transforms of the signals shown in Fig. P3.3-2b, c, d, e, and f. *Hint*: Time inversion in $g(t)$

results in the pulse $g_1(t)$ in Fig. P3.3-2b; consequently $g_1(t) = g(-t)$. The pulse in Fig. P3.3-2c can be expressed as $g(t-T) + g_1(t-T)$ [the sum of $g(t)$ and $g_1(t)$ both delayed by T]. The pulses in Fig. P3.3-2d and e both can be expressed as $g(t-T) + g_1(t+T)$ [the sum of $g(t)$ delayed by T and $g_1(t)$ advanced by T] for some suitable choice of T. The pulse in Fig. P3.3-2f can be obtained by time-expanding $g(t)$ by a factor of 2 and then delaying the resulting pulse by 2 seconds [or by first delaying $g(t)$ by 1 second and then time-expanding by a factor of 2].

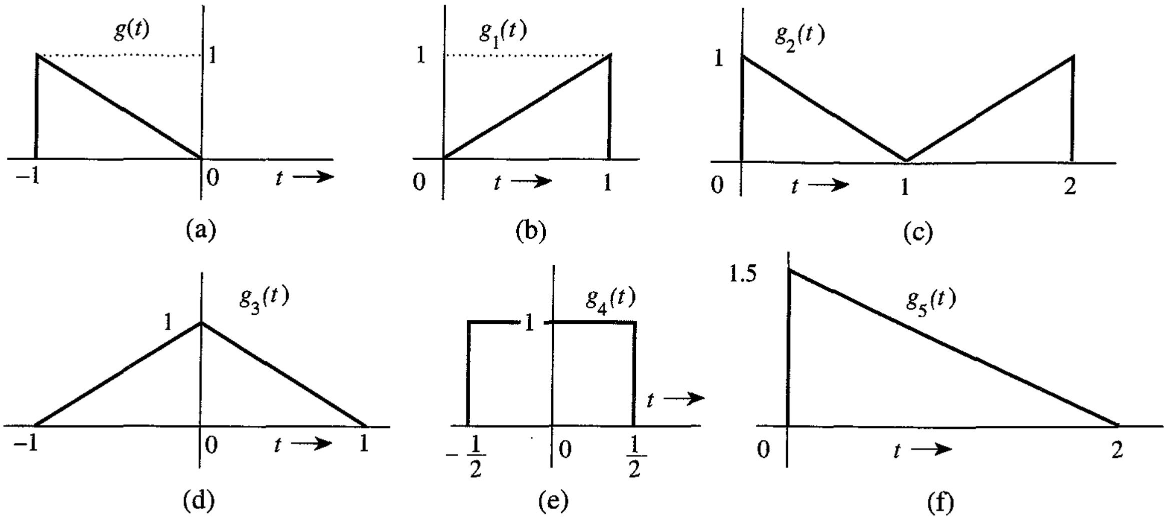


Figure P3.3-2

3.3-3 Using only the time-shifting property and Table 3.1, find the Fourier transforms of the signals shown in Fig. P3.3-3. *Hint*: The signal in Fig. P3.3-3a is a sum of two shifted gate pulses. The signal in Fig. P3.3-3b is $\sin t\,[u(t) - u(t-\pi)] = \sin t\, u(t) - \sin t\, u(t-\pi) = \sin t\, u(t) + \sin(t-\pi)\, u(t-\pi)$. The reader should verify that the addition of these two sinusoids indeed results in the pulse in Fig. P3.3-3b. In the same way we can express the signal in Figs. P3.3-3c as $\cos t\, u(t) + \sin(t-\pi/2)u(t-\pi/2)$ (verify this by sketching these signals). The signal in Fig. P3.3-3d is $e^{-at}[u(t) - u(t-T)] = e^{-at}u(t) - e^{-aT}e^{-a(t-T)}u(t-T)$.

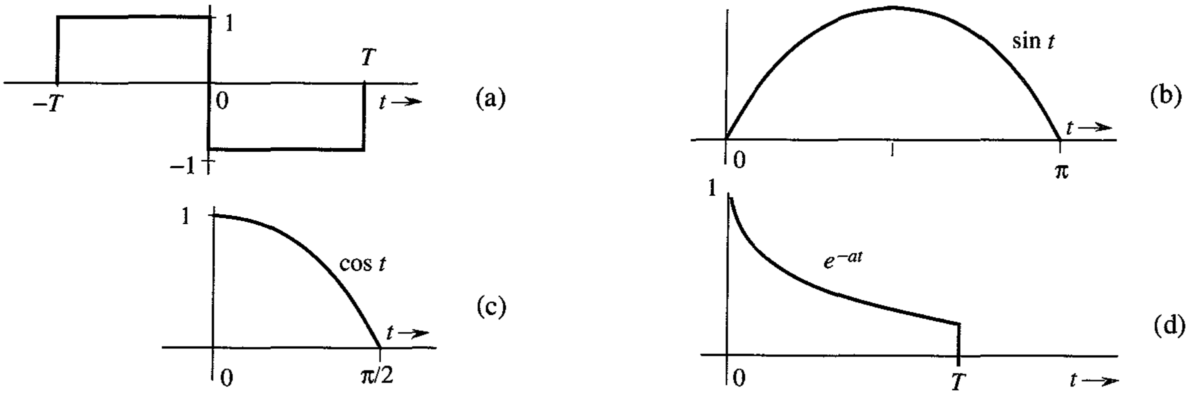


Figure P3.3-3

3.3-4 Using the time-shifting property, show that if $g(t) \Longleftrightarrow G(\omega)$, then

$$g(t+T) + g(t-T) \Longleftrightarrow 2G(\omega)\cos T\omega$$

This is the dual of Eq. (3.35). Using this result and pairs 17 and 19 in Table 3.1, find the Fourier transforms of the signals shown in Fig. P3.3-4.

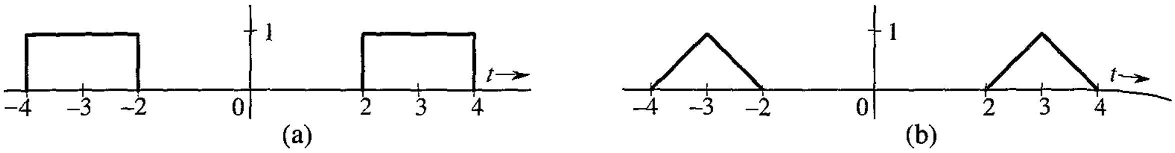


Figure P3.3-4

3.3-5 Prove the following results:

$$g(t)\sin\omega_0 t \Longleftrightarrow \frac{1}{2j}[G(\omega-\omega_0)-G(\omega+\omega_0)]$$

$$\frac{1}{2j}[g(t+T)-g(t-T)] \Longleftrightarrow G(\omega)\sin T\omega$$

Using the latter result and Table 3.1, find the Fourier transform of the signal in Fig. P3.3-5.

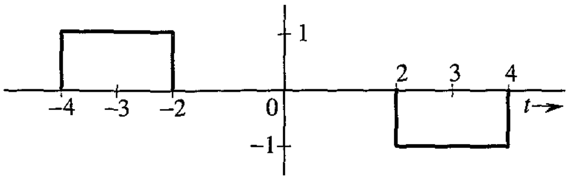


Figure P3.3-5

3.3-6 The signals in Fig. P3.3-6 are modulated signals with carrier cos $10t$. Find the Fourier transforms of these signals using the appropriate properties of the Fourier transform and Table 3.1. Sketch the amplitude and phase spectra for parts **(a)** and **(b)**. *Hint*: These functions can be expressed in the form $g(t)\cos\omega_0 t$.

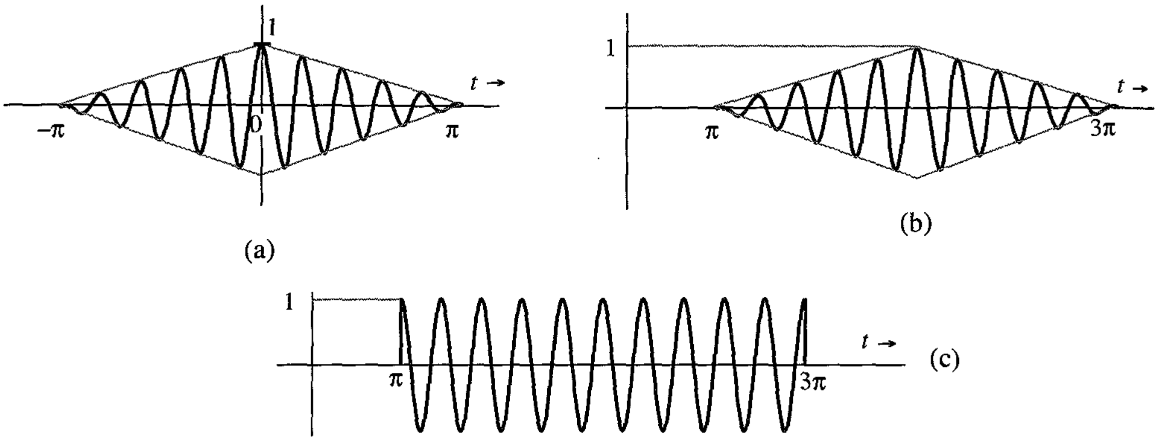


Figure P3.3-6

3.3-7 Using the frequency-shifting property and Table 3.1, find the inverse Fourier transform of the spectra shown in Fig. P3.3-7.

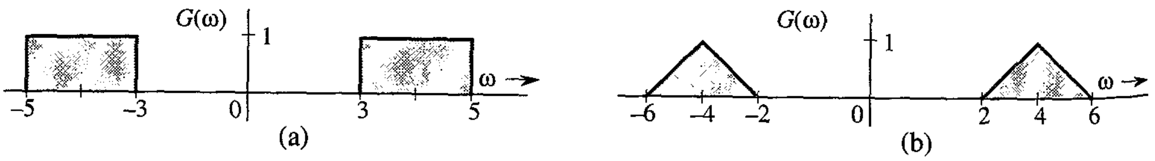


Figure P3.3-7

3.3-8 A signal $g(t)$ is band-limited to B Hz. Show that the signal $g^n(t)$ is band-limited to nB Hz. Hint: $g^2(t) \Longleftrightarrow [G(\omega) * G(\omega)]/2\pi$, and so on. Use the width property of convolution.

3.3-9 Find the Fourier transform of the signal in Fig. P3.3-3a by three different methods:

(a) By direct integration using the definition (3.8a).

(b) Using only pair 17 Table 3.1 and the time-shifting property.

(c) Using the time-differentiation and time-shifting properties, along with the fact that $\delta(t) \Longleftrightarrow 1$. Hint: $1 - \cos 2x = 2\sin^2 x$.

3.3-10 The process of recovering a signal $g(t)$ from the modulated signal $g(t)\cos \omega_0 t$ is called **demodulation**. Show that the signal $g(t)\cos \omega_0 t$ can be demodulated by multiplying it with $2\cos \omega_0 t$ and passing the product through a low-pass filter of bandwidth W rad/s [the bandwidth of $g(t)$]. Assume $W < \omega_0$. *Hint*: $2\cos^2 \omega_0 t = 1 + \cos 2\omega_0 t$. Recognize that the spectrum of $g(t)\cos 2\omega_0 t$ is centered at $2\omega_0$ and will be suppressed by a low-pass filter of bandwidth W rad/s.

3.4-1 Signals $g_1(t) = 10^4\text{rect}\,(10^4 t)$ and $g_2(t) = \delta(t)$ are applied at the inputs of the ideal low-pass filters $H_1(\omega) = \text{rect}\,(\omega/40{,}000\pi)$ and $H_2(\omega) = \text{rect}\,(\omega/20{,}000\pi)$ (Fig. P3.4-1). The outputs $y_1(t)$ and $y_2(t)$ of these filters are multiplied to obtain the signal $y(t) = y_1(t)y_2(t)$.

(a) Sketch $G_1(\omega)$ and $G_2(\omega)$.

(b) Sketch $H_1(\omega)$ and $H_2(\omega)$.

(c) Sketch $Y_1(\omega)$ and $Y_2(\omega)$.

(d) Find the bandwidths of $y_1(t)$, $y_2(t)$, and $y(t)$.

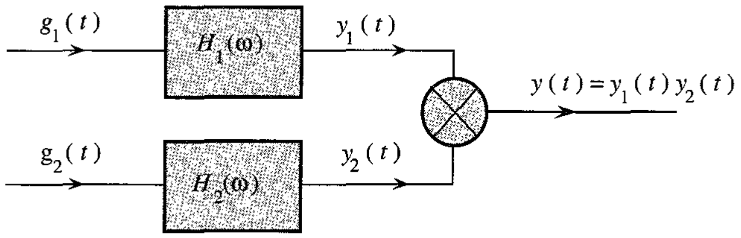


Figure P3.4-1

3.5-1 Consider a filter with the transfer function

$$H(\omega) = e^{-(k\omega^2 + j\omega t_0)}$$

Show that this filter is physically unrealizable by using the time-domain criterion [noncausal $h(t)$] and the frequency-domain (Paley-Wiener) criterion. Can this filter be made approximately realizable by choosing a sufficiently large t_0? Use your own (reasonable) criterion of approximate realizability to determine t_0. *Hint*: Use pair 22 in Table 3.1.

3.5-2 Show that a filter with transfer function

$$H(\omega) = \frac{2(10^5)}{\omega^2 + 10^{10}} e^{-j\omega t_0}$$

is unrealizable. Can this filter be made approximately realizable by choosing a sufficiently large t_0? Use your own (reasonable) criterion of approximate realizability to determine t_0. *Hint*: Show that the impulse response is noncausal.

3.5-3 Determine the maximum bandwidth of a signal that can be transmitted through the low-pass RC filter in Fig. 3.27a with $R = 1000$ and $C = 10^{-9}$ if, over this bandwidth, the amplitude response (gain) variation is to be within 5% and the time delay variation is to be within 2%.

3.5-4 A bandpass signal $g(t)$ of bandwidth $\Delta\omega = 2000$ centered at $\omega = 10^5$ is passed through the RC filter in Example 3.16 (Fig. 3.27a) with $RC = 10^{-3}$. If over the passband, the variation of less than 2% in amplitude response and less than 1% in time delay is considered distortionless transmission, would $g(t)$ be transmitted without distortion? Find the approximate expression for the output signal.

3.6-1 A certain channel has ideal amplitude, but nonideal phase response (Fig. P3.6-1), given by

$$|H(\omega)| = 1$$
$$\theta_h(\omega) = -\omega t_0 - k \sin \omega T \qquad k \ll 1$$

(a) Show that $y(t)$, the channel response to an input pulse $g(t)$ band-limited to B Hz, is

$$y(t) = g(t - t_0) + \frac{k}{2}[g(t - t_0 - T) - g(t - t_0 + T)]$$

Hint: Use $e^{-jk \sin \omega T} \approx 1 - jk \sin \omega T$.

(b) Discuss how this channel will affect TDM and FDM systems from the viewpoint of interference among the multiplexed signals.

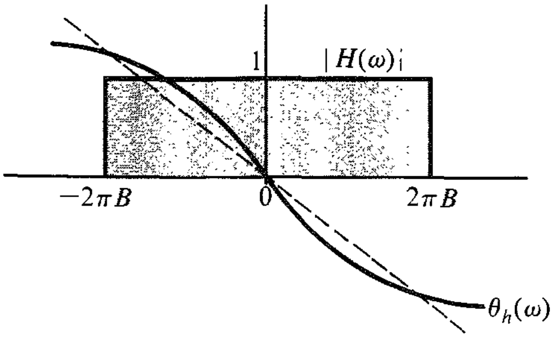


Figure P3.6-1

3.6-2 The distortion caused by multipath transmission can be partly corrected by a tapped delay equalizer. Show that if $\alpha \ll 1$, the distortion in the multipath system in Fig. 3.35a can be approximately corrected if the received signal in Fig. 3.35a is passed through the tapped delay equalizer shown in Fig. P3.6-2. *Hint*: From Eq. (3.63a), it is clear that the equalizer filter transfer function should be $H_{\text{eq}}(\omega) = 1/(1+\alpha\, e^{-j\omega\Delta t})$. Use the fact that $1/(1-x) = 1+x+x^2+x^3+\cdots$ if $x \ll 1$.

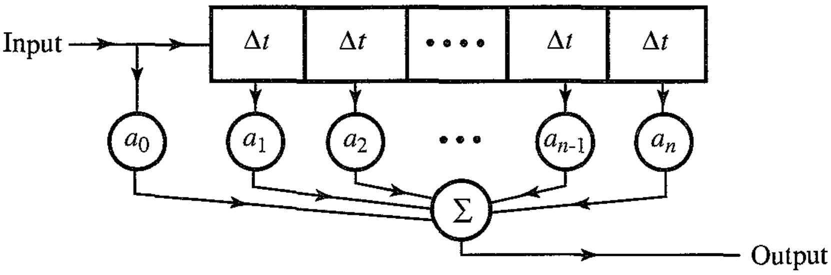


Figure P3.6-2

3.7-1 Show that the energy of the gaussian pulse

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{t^2}{2\sigma^2}}$$

is $1/2\sigma\sqrt{\pi}$. Verify this result by deriving the energy E_g from $G(\omega)$ using Parseval's theorem. *Hint*: See pair 22 in Table 3.1. Use the fact that

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

3.7-2 Show that

$$\int_{-\infty}^{\infty} \text{sinc}^2\,(kx)\,dx = \frac{\pi}{k}$$

Hint: Recognize that the integral is the energy of $g(t) = \text{sinc}\,(kt)$. Find this energy by using Parseval's theorem.

3.7-3 Generalize Parseval's theorem to show that for real, Fourier transformable signals $g_1(t)$ and $g_2(t)$

$$\int_{-\infty}^{\infty} g_1(t)g_2(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_1(-\omega)G_2(\omega)\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_1(\omega)G_2(-\omega)\,d\omega$$

3.7-4 Show that

$$\int_{-\infty}^{\infty} \text{sinc}\,(2\pi Bt - m\pi)\,\text{sinc}\,(2\pi Bt - n\pi)\,dt = \begin{cases} 0 & m \neq n \\ \dfrac{1}{2B} & m = n \end{cases}$$

Hint: Recognize that

$$\text{sinc}\,(2\pi Bt - k\pi) = \text{sinc}\left[2\pi B\left(t - \frac{k}{2B}\right)\right] \Longleftrightarrow \frac{1}{2B}\,\text{rect}\left(\frac{\omega}{4\pi B}\right)e^{-j\omega k/2B}$$

Use this fact and the result in Prob. 3.7-3 to show that

$$\int_{-\infty}^{\infty} \text{sinc}\,(2\pi Bt - m\pi)\,\text{sinc}\,(2\pi Bt - n\pi)\,dt = \frac{1}{8\pi B^2}\int_{-2\pi B}^{2\pi B} e^{j[(n-m)/2B]\omega}\,d\omega$$

The desired result follows from this integral.

3.7-5 For the signal

$$g(t) = \frac{2a}{t^2 + a^2}$$

determine the essential bandwidth B Hz of $g(t)$ such that the energy contained in the spectral components of $g(t)$ of frequencies below B Hz is 99% of the signal energy E_g. Hint: Determine $G(\omega)$ by applying the symmetry property [Eq. (3.24)] to pair 3 of Table 3.1.

3.7-6 A low-pass signal $g(t)$ is applied to a squaring device. The squarer output $g^2(t)$ is applied to a unity gain ideal low-pass filter of bandwidth Δf Hz (Fig. P3.7-6). Show that if Δf is very small ($\Delta f \to 0$), the filter output is a dc signal of amplitude $2E_g\Delta f$, where E_g is the energy of $g(t)$. *Hint*: The output $y(t)$ is a dc signal because its spectrum $Y(\omega)$ is concentrated at $\omega = 0$ from $-\Delta\omega$ to $\Delta\omega$ with $\Delta\omega \to 0$ (impulse at the origin). If $g^2(t) \Longleftrightarrow A(\omega)$, and $y(t) \Longleftrightarrow Y(\omega)$, then $Y(\omega) \approx [4\pi A(0)\Delta f]\delta(\omega)$.

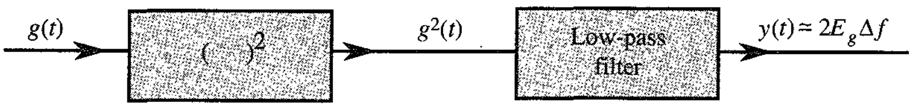


Figure P3.7-6

3.8-1 Show that the autocorrelation function of $g(t) = C \cos(\omega_0 t + \theta_0)$ is given by $\mathcal{R}_g(\tau) = (C^2/2) \cos \omega_0 \tau$, and the corresponding PSD is $S_g(\omega) = (C^2\pi/2)[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]$. Hence, show that for a signal $y(t)$ given by

$$y(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

the autocorrelation function and the PSD are given by

$$\mathcal{R}_y(\tau) = C_0{}^2 + \frac{1}{2}\sum_{n=1}^{\infty} C_n{}^2 \cos n\omega_0\tau$$

$$S_y(\omega) = 2\pi C_0{}^2\delta(\omega) + \frac{\pi}{2}\sum_{n=1}^{\infty} C_n{}^2[\delta(\omega - n\omega_0) + \delta(\omega + n\omega_0)]$$

Hint: Show that if $g(t) = g_1(t) + g_2(t)$, then $\mathcal{R}_g(\tau) = \mathcal{R}_{g_1}(\tau) + \mathcal{R}_{g_2}(\tau) + \mathcal{R}_{g_1g_2}(\tau) + \mathcal{R}_{g_2g_1}(\tau)$, where $\mathcal{R}_{g_1g_2}(\tau) = \lim_{T\to\infty}(1/T)\int_{-T/2}^{T/2} g_1(t)g_2(t+\tau)\,dt$. If $g_1(t)$ and $g_2(t)$ represent any two of the infinite terms in $y(t)$, then show that $\mathcal{R}_{g_1g_2}(\tau) = \mathcal{R}_{g_2g_1}(\tau) = 0$. To show this use the fact that the area under any sinusoid over a very large time interval is at most equal to the area of the half-cycle of the sinusoid.

3.8-2 The random binary signal $x(t)$ shown in Fig. P3.8-2 transmits one digit every T_b seconds. A binary **1** is transmitted by a pulse $p(t)$ of width $T_b/2$ and amplitude A; a binary **0** is transmitted by no pulse. The digits **1** and **0** are equally likely and occur randomly. Determine the autocorrelation function $\mathcal{R}_x(\tau)$ and the PSD $S_x(\omega)$.

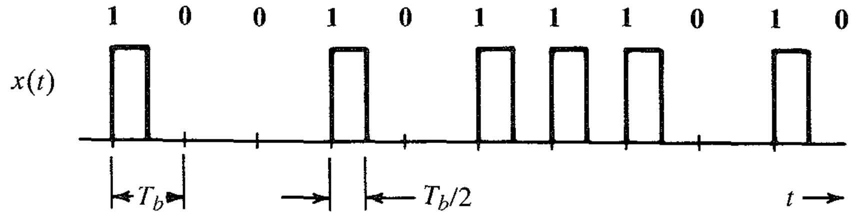


Figure P3.8-2

3.8-3 Find the mean square value (or power) of the output voltage $y(t)$ of the RC network shown in Fig. 3.27a with $RC = 1$ if the input voltage PSD $S_x(\omega)$ is given by: **(a)** K; **(b)** rect $(\omega/2)$; **(c)** $[\delta(\omega+1)+\delta(\omega-1)]$. In each case calculate the power (mean square value) of the input signal $x(t)$.

3.8-4 Find the mean square value (or power) of the output voltage $y(t)$ of the system shown in Fig. P3.8-4 if the input voltage PSD $S_x(\omega) = $ rect $(\omega/2)$. Calculate the power (mean square value) of the input signal $x(t)$.

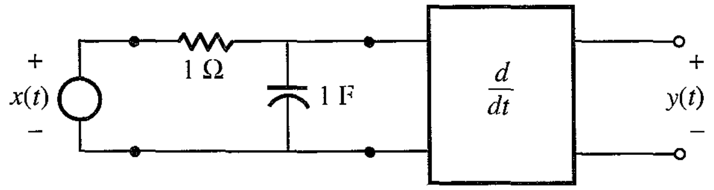


Figure P3.8-4

4 AMPLITUDE (LINEAR) MODULATION

Modulation is a process that causes a shift in the range of frequencies in a signal. It is used to gain certain advantages mentioned in Chapter 1. Before discussing modulation, it is important to distinguish between communication that does not use modulation (**baseband communication**) and communication that uses modulation (**carrier communication**).

4.1 BASEBAND AND CARRIER COMMUNICATION

The term **baseband** is used to designate the band of frequencies of the signal delivered by the source or the input transducer (see Fig. 1.2). In telephony, the baseband is the audio band (band of voice signals) of 0 to 3.5 kHz. In television, the baseband is the video band occupying 0 to 4.3 MHz. For digital data or PCM using bipolar signaling at a rate of R_b pulses per second, the baseband is 0 to R_b Hz.

In baseband communication, baseband signals are transmitted without modulation, that is, without any shift in the range of frequencies of the signal. Because the baseband signals have sizable power at low frequencies, they cannot be transmitted over a radio link but are suitable for transmission over a pair of wires, coaxial cables, or optical fibers. Local telephone communication, short-haul pulse-code modulation (PCM) (between two exchanges), and long-distance PCM over optical fibers use baseband communication. Modulation can be helpful in utilizing the vast spectrum of frequencies available because of technological advances. By modulating several baseband signals and shifting their spectra to nonoverlapping bands, one can use all the available bandwidth through frequency division multiplexing (FDM). Long-haul communication over a radio link also requires modulation to shift the signal spectrum to higher frequencies in order to enable efficient power radiation using antennas of reasonable dimensions. Yet another use of modulation is to exchange transmission bandwidth for the SNR.

Communication that uses modulation to shift the frequency spectrum of a signal is known as **carrier communication**. In this mode, one of the basic parameters (amplitude, frequency, or phase) of a **sinusoidal carrier** of high frequency ω_c is varied in proportion to the baseband signal $m(t)$. This results in amplitude modulation (AM), frequency modulation (FM), or phase modulation (PM), respectively. The latter two types of modulation are similar, and belong to the class of modulation known as **angle modulation**. Modulation is used to transmit analog as well as digital baseband signals.

A comment about pulse-modulated signals [pulse amplitude modulation (PAM), pulse width modulation (PWM), pulse position modulation (PPM), pulse code modulation (PCM), and delta modulation (DM)] is in order here. Despite the term modulation, these signals are baseband signals. The term modulation is used here in another sense. Pulse-modulation schemes are really baseband coding schemes, and they yield baseband signals. These signals must still modulate a carrier in order to shift their spectra.

4.2 AMPLITUDE MODULATION: DOUBLE SIDEBAND (DSB)

Amplitude modulation is characterized by the fact that the amplitude A of the **carrier** $A \cos(\omega_c t + \theta_c)$ is varied in proportion to the baseband (message) signal $m(t)$, the **modulating signal**. The frequency ω_c and the phase θ_c are constant. We can assume $\theta_c = 0$ without a loss of generality. If the carrier amplitude A is made directly proportional to the modulating signal $m(t)$, the **modulated signal** is $m(t) \cos \omega_c t$ (Fig. 4.1). As was seen earlier [Eq. (3.35)], this type of modulation simply shifts the spectrum of $m(t)$ to the carrier frequency (Fig. 4.1a). Thus, if

$$m(t) \Longleftrightarrow M(\omega)$$

then

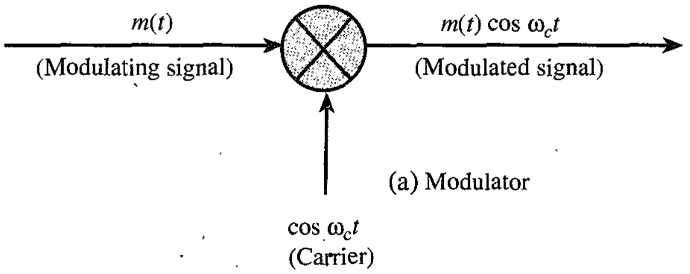
$$m(t) \cos \omega_c t \Longleftrightarrow \tfrac{1}{2}[M(\omega + \omega_c) + M(\omega - \omega_c)] \tag{4.1}$$

Recall that $M(\omega - \omega_c)$ is $M(\omega)$ shifted to the right by ω_c and $M(\omega + \omega_c)$ is $M(\omega)$ shifted to the left by ω_c. Thus, the process of modulation shifts the spectrum of the modulating signal to the left and the right by ω_c. Note also that if the bandwidth of $m(t)$ is B Hz, then, as seen from Fig. 4.1c, the bandwidth of the modulated signal is $2B$ Hz. We also observe that the modulated signal spectrum centered at ω_c is composed of two parts: a portion that lies above ω_c, known as the **upper sideband (USB)**, and a portion that lies below ω_c, known as the **lower sideband (LSB)**. Similarly, the spectrum centered at $-\omega_c$ has upper and lower sidebands. Hence, this is a modulation scheme with double sidebands. We shall see a little later that the modulated signal in this scheme does not contain a discrete component of the carrier frequency ω_c. For this reason it is called **double-sideband suppressed carrier (DSB-SC) modulation.**

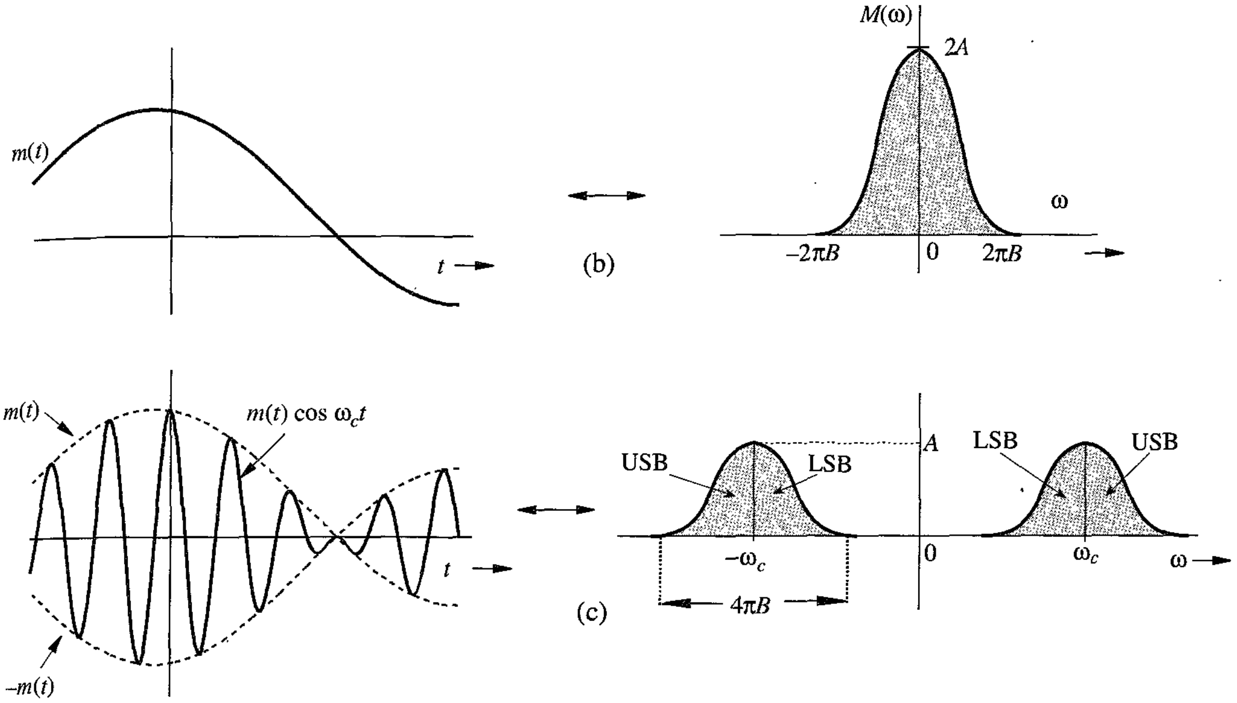
The relationship of B to ω_c is of interest. Figure 4.1c shows that $\omega_c \geq 2\pi B$ in order to avoid the overlap of the spectra centered at ω_c and $-\omega_c$. If $\omega_c < 2\pi B$, these spectra overlap and the information of $m(t)$ is lost in the process of modulation, which makes it impossible to get back $m(t)$ from the modulated signal $m(t) \cos \omega_c t$.*

---

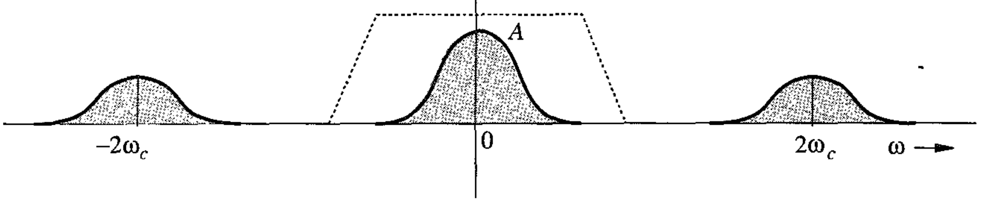
* Practical factors may impose additional restrictions on ω_c. For instance, in the case of broadcast applications, a radiating antenna can radiate only a narrow band without distortion. This means that to avoid distortion caused by the radiating antenna, $\omega_c/2\pi B \gg 1$. The broadcast band AM radio, for instance, with $B = 5$ kHz and the band of 550 to 1600 kHz for the carrier frequency give a ratio of $\omega_c/2\pi B$ roughly in the range of 100 to 300.

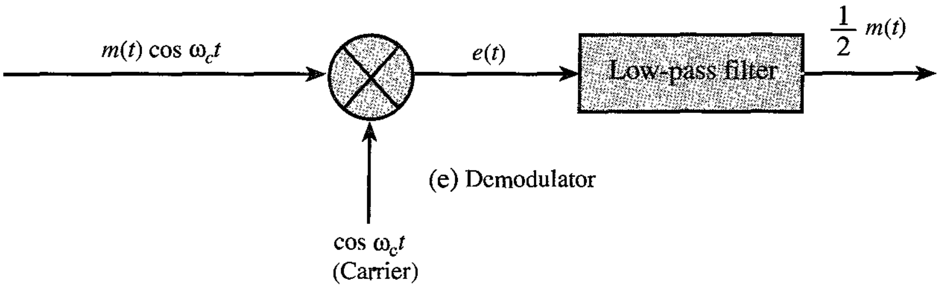

$m(t)$
(Modulating signal)
$m(t) \cos \omega_c t$
(Modulated signal)
$\cos \omega_c t$
(Carrier)

(a) Modulator



(d) Spectrum of $e(t)$




$m(t) \cos \omega_c t$
$e(t)$
Low-pass filter
$\frac{1}{2} m(t)$
$\cos \omega_c t$
(Carrier)

(e) Demodulator

Figure 4.1 DSB-SC modulation and demodulation.

Demodulation

The DSB-SC modulation translates or shifts the frequency spectrum to the left and the right by ω_c (that is, at $+\omega_c$ and $-\omega_c$), as seen from Eq. (4.1). To recover the original signal $m(t)$ from the modulated signal, it is necessary to retranslate the spectrum to its original position. The process of recovering the signal from the modulated signal (retranslating the spectrum to its original position) is referred to as **demodulation**, or **detection**. Observe that if the modulated signal spectrum in Fig. 4.1c is shifted to the left and to the right by ω_c (and multiplied by one-half), we obtain the spectrum shown in Fig. 4.1d, which contains the desired baseband spectrum plus an unwanted spectrum at $\pm 2\omega_c$. The latter can be suppressed by a low pass filter. Thus, demodulation, which is almost identical to modulation, consists of multiplication of the incoming modulated signal $m(t)\cos\omega_c t$ by a carrier $\cos\omega_c t$ followed by a low pass filter, as shown in Fig. 4.1e. We can verify this conclusion directly in the time domain by observing that the signal $e(t)$ in Fig. 4.1e is

$$\begin{aligned} e(t) &= m(t)\cos^2\omega_c t \\ &= \frac{1}{2}[m(t) + m(t)\cos 2\omega_c t] \end{aligned} \tag{4.2a}$$

Therefore, the Fourier transform of the signal $e(t)$ is

$$E(\omega) = \frac{1}{2}M(\omega) + \frac{1}{4}[M(\omega + 2\omega_c) + M(\omega - 2\omega_c)] \tag{4.2b}$$

This shows that the signal $e(t)$ consists of two components $(1/2)m(t)$ and $(1/2)m(t)\cos 2\omega_c t$, with their spectra as shown in Fig. 4.1d. The spectrum of the second component, being a modulated signal with carrier frequency $2\omega_c$, is centered at $\pm 2\omega_c$. Hence, this component is suppressed by the low pass filter in Fig. 4.1e. The desired component $(1/2)M(\omega)$, being a low pass spectrum (centered at $\omega = 0$), passes through the filter unharmed, resulting in the output $(1/2)m(t)$. We can get rid of the inconvenient fraction 1/2 in the output by using a carrier $2\cos\omega_c t$ instead of $\cos\omega_c t$. In fact, in future, we shall often use this strategy, which does not affect general conclusions.

A possible form of low pass filter characteristics is shown (dotted) in Fig. 4.1d. This method of recovering the baseband signal is called **synchronous detection**, or **coherent detection**, where we use a carrier of exactly the same frequency (and phase) as the carrier used for modulation. Thus, for demodulation, we need to generate a local carrier at the receiver in frequency and phase coherence (synchronism) with the carrier used at the modulator.

EXAMPLE 4.1 For a baseband signal $m(t) = \cos\omega_m t$, find the DSB-SC signal, and sketch its spectrum. Identify the USB and LSB. Verify that the DSB-SC modulated signal can be demodulated by the demodulator in Fig. 4.1e.

The case in this example is referred to as **tone modulation** because the modulating signal is a pure sinusoid, or tone, $\cos\omega_m t$. We shall work this problem in the frequency domain as well as the time domain in order to clarify the basic concepts of DSB-SC modulation. In the frequency domain approach, we work with the signal spectra. The spectrum of the baseband signal $m(t) = \cos\omega_m t$ is given by

$$M(\omega) = \pi[\delta(\omega - \omega_m) + \delta(\omega + \omega_m)]$$

The spectrum consists of two impulses located at $\pm\omega_m$, as shown in Fig. 4.2a. The DSB-SC (modulated) spectrum, as seen from Eq. (4.1), is the baseband spectrum in Fig. 4.2*a* shifted to the right and the left by ω_c (times one-half), as shown in Fig. 4.2b. This spectrum consists of impulses at $\pm(\omega_c - \omega_m)$ and $\pm(\omega_c + \omega_m)$. The spectrum beyond ω_c is the USB, and the one below ω_c is the LSB. Observe that the DSB-SC spectrum does not have the component of the carrier frequency ω_c. This is why it is called **suppressed carrier**.

In the time-domain approach, we work directly with signals in the time domain. For the baseband signal $m(t) = \cos \omega_m t$, the DSB-SC signal $\varphi_{\text{DSB-SC}}(t)$ is

$$\begin{aligned}\varphi_{\text{DSB-SC}}(t) &= m(t) \cos \omega_c t \\ &= \cos \omega_m t \cos \omega_c t \\ &= \frac{1}{2}[\cos (\omega_c + \omega_m)t + \cos (\omega_c - \omega_m)t]\end{aligned}$$

This shows that when the baseband (message) signal is a single sinusoid of frequency ω_m, the modulated signal consists of two sinusoids: the component of frequency $\omega_c + \omega_m$ (the USB) and the component of frequency $\omega_c - \omega_m$ (the LSB). Figure 4.2b shows precisely the spectrum of $\varphi_{\text{DSB-SC}}(t)$. Thus, each component of frequency ω_m in the modulating signal results into two components of frequencies $\omega_c + \omega_m$ and $\omega_c - \omega_m$ in the modulated signal. Note the curious fact that there is no component of the carrier frequency ω_c on the right-hand side of the preceding equation. As mentioned, this is why it is called double sideband-suppressed carrier (DSB-SC) modulation.*

We now verify that the modulated signal $\varphi_{\text{DSB-SC}}(t) = \cos \omega_m t \cos \omega_c t$, when applied to the input of the demodulator in Fig. 4.1e, yields the output proportional to the desired baseband signal $\cos \omega_m t$. The signal $e(t)$ in Fig. 4.1e is given by

$$\begin{aligned}e(t) &= \cos \omega_m t \cos^2 \omega_c t \\ &= \frac{1}{2} \cos \omega_m t \, (1 + \cos 2\omega_c t)\end{aligned}$$

The spectrum of the term $\cos \omega_m t \cos 2\omega_c t$ is centered at $2\omega_c$, and will be suppressed by the low-pass filter, yielding $\frac{1}{2} \cos \omega_m t$ as the output. We can also derive this result in the frequency domain. Demodulation causes the spectrum in Fig. 4.2b to shift left and right by ω_c (and multiplies by one-half). This results in the spectrum shown in Fig. 4.2c. The low-pass filter suppresses the spectrum centered at $\pm 2\omega_c$, yielding the spectrum $\frac{1}{2}M(\omega)$.

Modulators

Modulation can be achieved in several ways. We shall discuss here some important categories of modulators.

Multiplier Modulators: Here modulation is achieved directly by multiplying $m(t)$ by $\cos \omega_c t$ using an analog multiplier whose output is proportional to the product of two input

---

* The term suppressed carrier does not necessarily mean absence of the spectrum at the carrier frequency. It means that there is no discrete component of the carrier frequency. This implies that the spectrum of the DSB-SC does not have impuses at $\pm\omega_c$, which also implies that the modulated signal $m(t) \cos \omega_c t$ does not contain a term of the form $k \cos \omega_c t$ [assuming that $m(t)$ has a zero mean value].

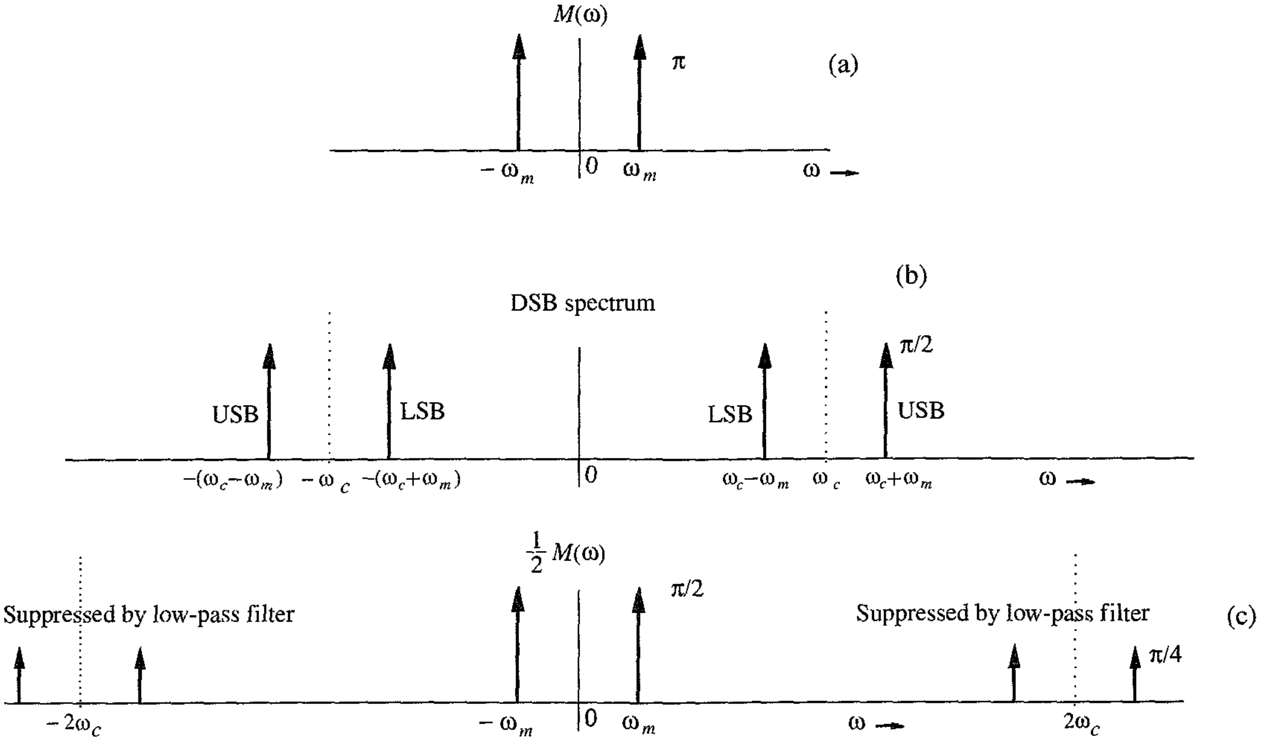


Figure 4.2 Example of DSB-SC modulation.

signals.* It is rather difficult to maintain linearity in this kind of amplifier, and they tend to be rather expensive. It is best to avoid them if possible. For practical implementation of such modulators, see Sheingold.[1]

Nonlinear Modulators: Modulation can also be achieved by using nonlinear devices, such as a semiconductor diode or a transistor. Figure 4.3 shows one possible scheme, which uses two identical nonlinear elements shown by boxes marked NL.

Let the input-output characteristics of either of the nonlinear elements be approximated by a power series:

$$y(t) = ax(t) + bx^2(t) \tag{4.3}$$

where $x(t)$ and $y(t)$ are the input and the output, respectively, of the nonlinear element. The summer output $z(t)$ in Fig. 4.3 is given by

$$z(t) = y_1(t) - y_2(t) = [ax_1(t) + bx_1^2(t)] - [ax_2(t) + bx_2^2(t)]$$

Substituting the two inputs $x_1(t) = \cos\omega_c t + m(t)$ and $x_2(t) = \cos\omega_c t - m(t)$ in this equation yields

$$z(t) = 2am(t) + 4bm(t)\,\cos\omega_c t$$

---

* Such a multiplier may be obtained from a variable-gain amplifier in which the gain parameter (such as the β of a transistor) is controlled by one of the signals, say, $m(t)$. When the signal $\cos\omega_c t$ is applied at the input of this amplifier, the output is proportional to $m(t)\cos\omega_c t$.

Another way to multiply two signals is through logarithmic amplifiers. Here, the basic components are a logarithmic and an antilogarithmic amplifier with outputs proportional to the log and antilog of their inputs, respectively. Using two logarithmic amplifiers, we generate and add the logarithms of the two signals to be multiplied. The sum is then applied to an antilogarithmic amplifier to obtain the desired product.

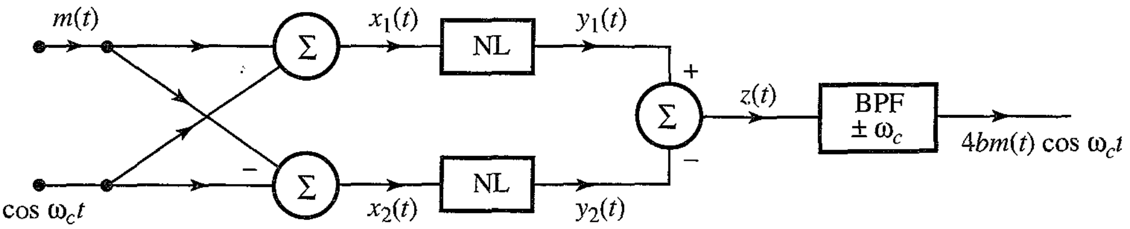


Figure 4.3 Nonlinear DSB-SC modulator.

The spectrum of $m(t)$ is centered at the origin, whereas the spectrum of $m(t)\cos\omega_c t$ is centered at $\pm\omega_c$. Consequently, when $z(t)$ is passed through a bandpass filter tuned to ω_c, the signal $am(t)$ is suppressed and the desired modulated signal $4bm(t)\cos\omega_c t$ passes through unharmed.

In this circuit there are two inputs: $m(t)$ and $\cos\omega_c t$. The summer output $z(t)$ does not contain one of the inputs, the carrier signal $\cos\omega_c t$. Consequently, the carrier signal does not appear at the input of the final bandpass filter. The circuit acts as a balanced bridge for one of the inputs (the carrier). Circuits which have this characteristic are called **balanced circuits**. The nonlinear modulator in Fig. 4.3 is an example of a class of modulators known as **balanced modulators**. This circuit is balanced with respect to only one input (the carrier); the other input $m(t)$ still appears at the final bandpass filter, which must reject it. For this reason, it is called a **single balanced modulator**. A circuit balanced with respect to both inputs is called a **double balanced modulator**, of which the ring modulator (see Fig. 4.6) is an example.

Switching Modulators: The multiplication operation required for modulation can be replaced by a simpler switching operation if we realize that a modulated signal can be obtained by multiplying $m(t)$ not only by a pure sinusoid but by any periodic signal $\phi(t)$ of the fundamental radian frequency ω_c. Such a periodic signal can be expressed by a trigonometric Fourier series as

$$\phi(t) = \sum_{n=0}^{\infty} C_n \cos(n\omega_c t + \theta_n) \tag{4.4a}$$

Hence,

$$m(t)\phi(t) = \sum_{n=0}^{\infty} C_n\, m(t) \cos(n\omega_c t + \theta_n) \tag{4.4b}$$

This shows that the spectrum of the product $m(t)\phi(t)$ is the spectrum $M(\omega)$ shifted to $\pm\omega_c,\ \pm 2\omega_c,\ \ldots,\ \pm n\omega_c, \ldots$. If this signal is passed through a bandpass filter of bandwidth $2B$ Hz and tuned to ω_c, then we get the desired modulated signal $c_1 m(t)\cos(\omega_c t + \theta_1)$.*

The square pulse train $w(t)$ in Fig. 4.4b is a periodic signal whose Fourier series was found earlier [Eq. (2.75)] as

$$w(t) = \frac{1}{2} + \frac{2}{\pi}\left(\cos\omega_c t - \frac{1}{3}\cos 3\omega_c t + \frac{1}{5}\cos 5\omega_c t - \cdots\right) \tag{4.5}$$

The signal $m(t)w(t)$ is given by

$$m(t)w(t) = \frac{1}{2}m(t) + \frac{2}{\pi}\left[m(t)\cos\omega_c t - \frac{1}{3}m(t)\cos 3\omega_c t + \frac{1}{5}m(t)\cos 5\omega_c t - \cdots\right] \tag{4.6}$$

---

* The phase θ_1 is not important.

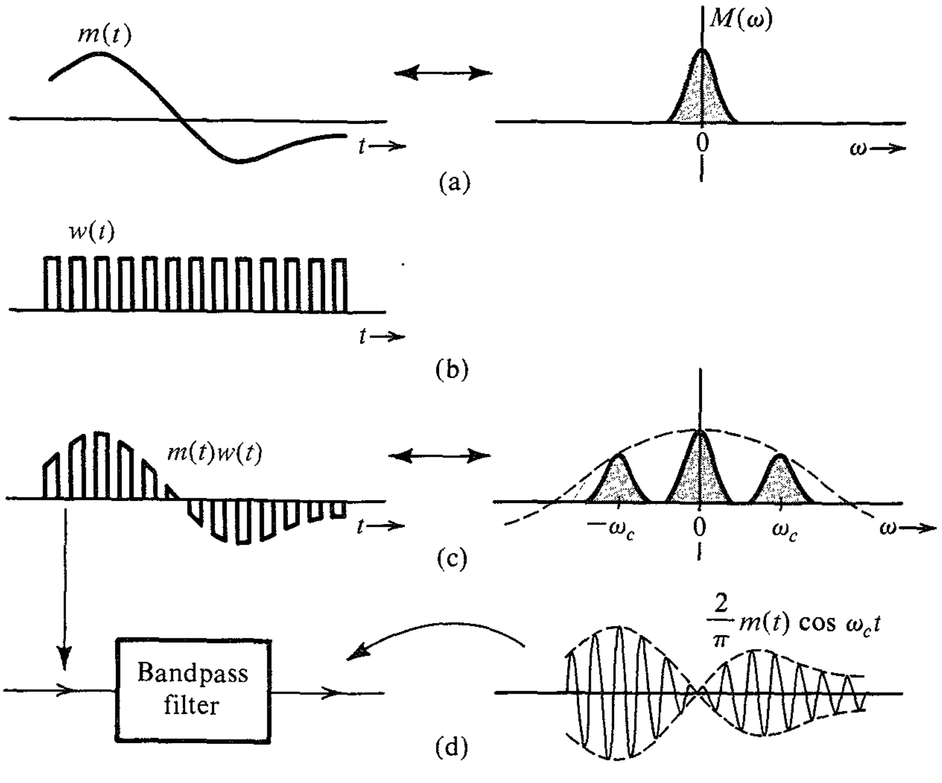


Figure 4.4 Switching modulator for DSB-SC.

The signal $m(t)w(t)$ consists not only of the component $m(t)$ but also of an infinite number of modulated signals with carrier frequencies ω_c, $3\omega_c$, $5\omega_c$, $\cdots$. Therefore, the spectrum of $m(t)w(t)$ consists of the spectrum $M(\omega)$ and $M(\omega)$ shifted to $\pm\omega_c$, $\pm 3\omega_c$, $\pm 5\omega_c$, $\cdots$ (with decreasing relative weights), as shown in Fig. 4.4c. We are interested in the modulated component $m(t)\cos\omega_c t$ only. To separate this component from the rest of the crowd, we pass the signal $m(t)w(t)$ through a bandpass filter of bandwidth $2B$ Hz, centered at the frequency $\pm\omega_c$. This will suppress all the spectral components not centered at $\pm\omega_c$ to yield the desired modulated signal $(2/\pi)m(t)\cos\omega_c t$ (Fig. 4.4d).

We now see the real payoff of this method. Multiplication of a signal by a square pulse train is in reality a switching operation. It involves switching the signal $m(t)$ on and off periodically and can be accomplished by simple switching elements controlled by $w(t)$. Figure 4.5a shows one such electronic switch, the **diode-bridge modulator**, driven by a sinusoid $A\cos\omega_c t$ to produce the switching action. Diodes D_1, D_2 and D_3, D_4 are matched pairs. When the signal $\cos\omega_c t$ is of a polarity that will make terminal c positive with respect to d, all the diodes conduct. Because diodes D_1 and D_2 are matched, terminals a and b have the same potential and are effectively shorted. During the next half-cycle, terminal d is positive with respect to c, and all four diodes open, thus, opening the terminals a and b. The diode bridge in Fig. 4.5a, therefore, serves as a desired electronic switch, where the terminals a and b open and close periodically with the carrier frequency f_c when a sinusoid $A\cos\omega_c t$ is applied across the terminals cd. To obtain the signal $m(t)w(t)$, we may place this electronic switch (terminals ab) in series (Fig. 4.5b) or across (in parallel) $m(t)$, as shown in Fig. 4.5c. These modulators are known as the **series-bridge diode modulator** and the **shunt-bridge diode modulator**, respectively. This switching on and off of $m(t)$ repeats for each cycle of the carrier, resulting in the switched signal $m(t)w(t)$, which when bandpass filtered, yields the desired modulated signal $(2/\pi)m(t)\cos\omega_c t$.

Another switching modulator, known as the **ring modulator**, is shown in Fig. 4.6a. During the positive half-cycles of the carrier, diodes D_1 and D_3 conduct, and D_2 and D_4 are open. Hence, terminal a is connected to c, and terminal b is connected to d. During the negative half-cycles of the carrier, diodes D_1 and D_3 are open, and D_2 and D_4 are conducting, thus connecting terminal a to d and terminal b to c. Hence, the output is proportional to $m(t)$ during the positive half-cycle and to $-m(t)$ during the negative half-cycle. In effect, $m(t)$ is multiplied by a square pulse train $w_0(t)$, shown in Fig. 4.6b. The Fourier series for $w_0(t)$ as found in Eq. (2.76) is

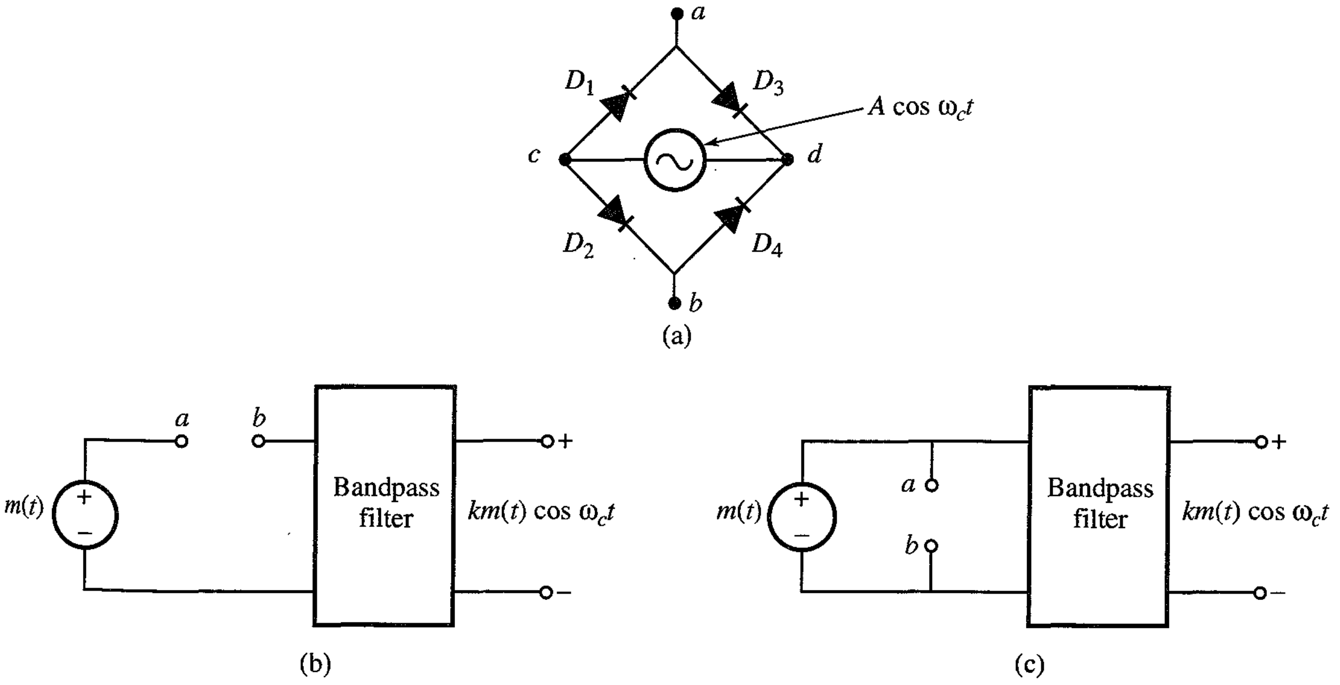
$$w_0(t) = \frac{4}{\pi}\left(\cos\omega_c t - \frac{1}{3}\cos 3\omega_c t + \frac{1}{5}\cos 5\omega_c t - \cdots\right) \tag{4.7a}$$

and

$$v_i(t) = m(t)w_0(t) = \frac{4}{\pi}\left[m(t)\cos\omega_c t - \frac{1}{3}m(t)\cos 3\omega_c t + \frac{1}{5}m(t)\cos 5\omega_c t - \cdots\right] \tag{4.7b}$$

The signal $m(t)w_0(t)$ is shown in Fig. 4.6d. When this waveform is passed through a bandpass filter tuned to ω_c (Fig. 4.6a), the filter output will be the desired signal $(4/\pi)m(t)\cos\omega_c t$.

In this circuit there are two inputs: $m(t)$ and $\cos\,\omega_c t$. The input to the final bandpass filter does not contain either of these inputs. Consequently, this circuit is an example of a **double balanced modulator**.

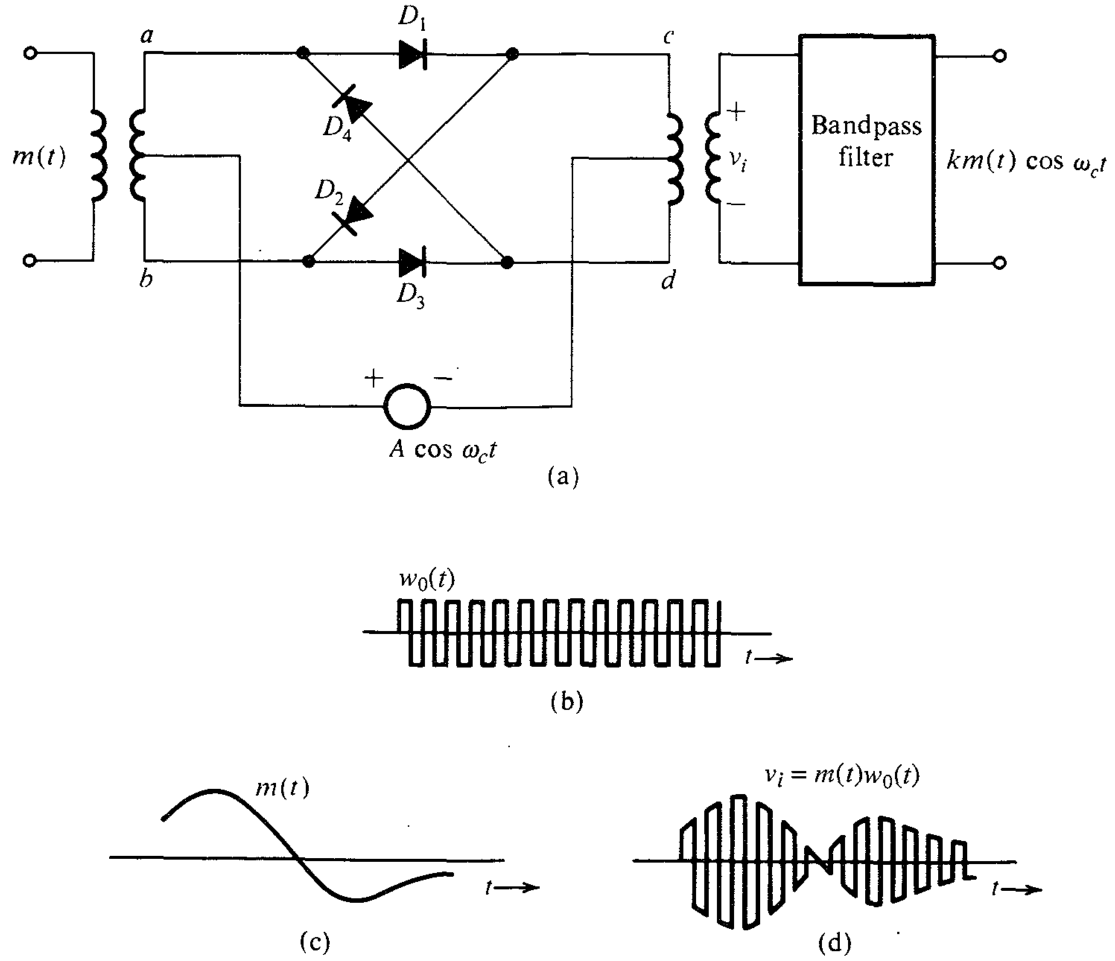


(a)

(b)

(c)

Figure 4.5 (a) Diode-bridge electronic switch. (b) Series-bridge diode modulator. (c) Shunt-bridge diode modulator.


D_1
a
c
$m(t)$
D_4
D_2
+
v_i
−
Bandpass filter
$km(t) \cos \omega_c t$
b
d
D_3
+
−
$A \cos \omega_c t$
$w_0(t)$
$t \rightarrow$
$m(t)$
$t \rightarrow$
$v_i = m(t)w_0(t)$
$t \rightarrow$

(a)

(b)

(c)

(d)

Figure 4.6 Ring modulator.

EXAMPLE 4.2 **Frequency Mixer or Converter**

We shall analyze a frequency mixer, or frequency converter, used to change the carrier frequency of a modulated signal $m(t) \cos \omega_c t$ from ω_c to some other frequency ω_I.

This can be done by multiplying $m(t) \cos \omega_c t$ by $2 \cos \omega_{\text{mix}} t$, where $\omega_{\text{mix}} = \omega_c + \omega_I$ or $\omega_c - \omega_I$, and then bandpass-filtering the product, as shown in Fig. 4.7a.

The product $x(t)$ is

$$x(t) = 2m(t) \cos \omega_c t \cos \omega_{\text{mix}} t$$

$$= m(t)[\cos (\omega_c - \omega_{\text{mix}})t + \cos (\omega_c + \omega_{\text{mix}})t]$$

If we select $\omega_{\text{mix}} = \omega_c - \omega_I$,

$$x(t) = m(t)[\cos \omega_I t + \cos (2\omega_c - \omega_I)t]$$

If we select $\omega_{\text{mix}} = \omega_c + \omega_I$,

$$x(t) = m(t)[\cos \omega_I t + \cos (2\omega_c + \omega_I)t]$$

In either case, a bandpass filter at the output, tuned to ω_I, will pass the term $m(t)$ cos $\omega_I t$ and suppress the other term, yielding the output $m(t)$ cos $\omega_I t$.* Thus, the carrier frequency has been translated to ω_I from ω_c.

The operation of frequency mixing, or frequency conversion (also known as heterodyning), is identical to the operation of modulation with a modulating carrier frequency (the mixer oscillator frequency ω_{mix}) that differs from the incoming carrier frequency by ω_I. Any one of the modulators discussed earlier can be used for frequency mixing. When we select the local carrier frequency $\omega_{\text{mix}} = \omega_c + \omega_I$, the operation is called **up-conversion**, and when we select $\omega_{\text{mix}} = \omega_c - \omega_I$, the operation is **down-conversion**.

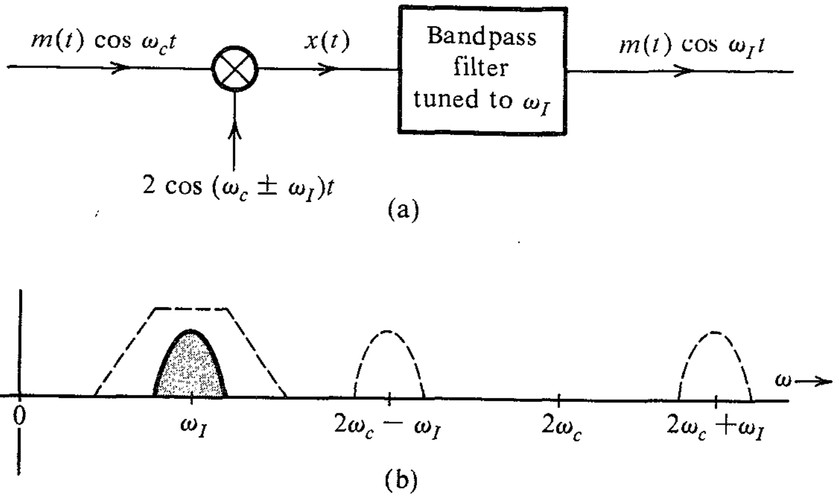


Figure 4.7 Frequency mixer or converter.

Demodulation of DSB-SC Signals

As discussed earlier, demodulation of a DSB-SC signal is identical to modulation (see Fig. 4.1). At the receiver, we multiply the incoming signal by a local carrier of frequency and phase in synchronism with the carrier used at the modulator. The product is then passed through a low-pass filter. The only difference between the modulator and the demodulator is the output filter. In the modulator, the multiplier output is passed through a bandpass filter tuned to ω_c, whereas in the demodulator, the multiplier output is passed through a low-pass filter. Therefore, all the modulators discussed earlier can also be used as demodulators, provided the bandpass filters at the output are replaced by low-pass filters of bandwidth B.

For demodulation, the receiver must generate a carrier in phase and frequency synchronism with the incoming carrier. These demodulators are called **synchronous** or **coherent** (also **homodyne**) demodulators.†

EXAMPLE 4.3 Analyze the switching demodulator that uses the electronic switch (diode bridge) in Fig. 4.5 as a switch (either in series or in parallel).

---

* Assuming that $\omega_c - \omega_I \geq 2\pi B$ and $\omega_I \geq 2\pi B$ so that various spectra in Fig. 4.7b do not overlap.

† The terms synchronous, coherent, and homodyne mean the same thing. The term homodyne is used in contrast to heterodyne where a different carrier frequency is used for the purpose of translating the spectrum (see Example 4.2).

The input signal is $m(t)\cos\omega_c t$. The carrier causes the periodic switching on and off of the input signal. Therefore, the output is $m(t)\cos\omega_c t \times w(t)$. Using the identity $\cos x \cos y = 0.5[\cos(x+y) + \cos(x-y)]$, we obtain

$$\begin{aligned} m(t)\cos\omega_c t \times w(t) &= m(t)\cos\omega_c t\left[\frac{1}{2} + \frac{2}{\pi}\left(\cos\omega_c t - \frac{1}{3}\cos 3\omega_c t + \cdots\right)\right] \\ &= \frac{2}{\pi}m(t)\cos^2\omega_c t + \text{terms of the form } m(t)\cos n\omega_c t \\ &= \frac{1}{\pi}m(t) + \frac{1}{\pi}m(t)\cos 2\omega_c t + \text{terms of the form } m(t)\cos n\omega_c t \end{aligned}$$

Spectra of the terms of the form $m(t)\cos n\omega_c t$ are centered at $\pm n\omega_c$ and are filtered out by the low-pass filter yielding the output $(1/\pi)m(t)$. It is left as an exercise for the reader to show that the output of the ring demodulator in Fig. 4.6a (with the low-pass filter at the output) is $(2/\pi)m(t)$ (twice that of the switching demodulator in this example).

4.3 AMPLITUDE MODULATION (AM)

For the suppressed carrier scheme discussed in the last section, a receiver must generate a carrier in frequency and phase synchronism with the carrier at the transmitter that may be located hundreds or thousands of miles away. This calls for a sophisticated receiver and could be quite costly. The other alternative is for the transmitter to transmit a carrier $A\cos\omega_c t$ [along with the modulated signal $m(t)\cos\omega_c t$] so that there is no need to generate a carrier at the receiver. In this case the transmitter needs to transmit much larger power, which makes it rather expensive. In point-to-point communications, where there is one transmitter for each receiver, substantial complexity in the receiver system can be justified, provided it results in a large enough saving in expensive high-power transmitting equipment. On the other hand, for a broadcast system with a multitude of receivers for each transmitter, it is more economical to have one expensive high-power transmitter and simpler, less expensive receivers. The second option (transmitting a carrier along with the modulated signal) is the obvious choice for this case. This is the so-called AM (amplitude modulation), in which the transmitted signal $\varphi_{\text{AM}}(t)$ is given by

$$\begin{aligned} \varphi_{\text{AM}}(t) &= A\cos\omega_c t + m(t)\cos\omega_c t && \text{(4.8a)} \\ &= [A + m(t)]\cos\omega_c t && \text{(4.8b)} \end{aligned}$$

The spectrum of $\varphi_{\text{AM}}(t)$ is the same as that of $m(t)\cos\omega_c t$ plus two additional impulses at $\pm\omega_c$,

$$\varphi_{\text{AM}}(t) \Longleftrightarrow \frac{1}{2}[M(\omega+\omega_c) + M(\omega-\omega_c)] + \pi A[\delta(\omega+\omega_c) + \delta(\omega-\omega_c)] \quad \text{(4.8c)}$$

Recall that the DSB-SC signal is $m(t)\cos\omega_c t$. From Eq. (4.8b) it follows that the AM signal is identical to the DSB-SC signal with $A + m(t)$ as the modulating signal [instead of $m(t)$]. Therefore, to sketch $\varphi_{\text{AM}}(t)$, we sketch $A + m(t)$ and $-[A + m(t)]$ and fill in between with the sinusoid of the carrier frequency. Two cases are considered in Fig. 4.8. In the first case, A is large enough so that $A + m(t) \geq 0$ (is nonnegative) for all values of t. In the second case, A is not large enough to satisfy this condition. In the first case, the envelope has the same shape as $m(t)$ (although riding on a dc of magnitude A). In the second case, the envelope shape is not $m(t)$ because some parts get rectified. This means we can detect the desired signal $m(t)$ by detecting the envelope in the first case. Such a detection is not possible in the second case. We shall see that the envelope detection is an extremely simple and inexpensive operation, which does not require generation of a local carrier for the demodulation. But as seen above the envelope of AM has the information about $m(t)$ only if the AM signal $[A + m(t)] \cos \omega_c t$ satisfies the condition $A + m(t) > 0$ for all t.

Recall also that the envelope of a signal $E(t) \cos \omega_c t$ is $E(t)$ provided $E(t) \geq 0$ for all t.* This means [see Eq. (4.8b)] that $A + m(t)$ is the envelope of $\varphi_{\text{AM}}(t)$ only if $A + m(t) \geq 0$ for all t. This conclusion is readily verified from Fig. 4.8d and e. In Fig. 4.8d, where $A + m(t) \geq 0$, $A + m(t)$ is indeed the envelope, and $m(t)$ can be recovered from this envelope. In Fig. 4.8e, where $A + m(t)$ is not always positive, the envelope is not $A + m(t)$, but rectified $A + m(t)$, and $m(t)$ cannot be recovered from the envelope. Consequently, demodulation of $\varphi_{\text{AM}}(t)$ in

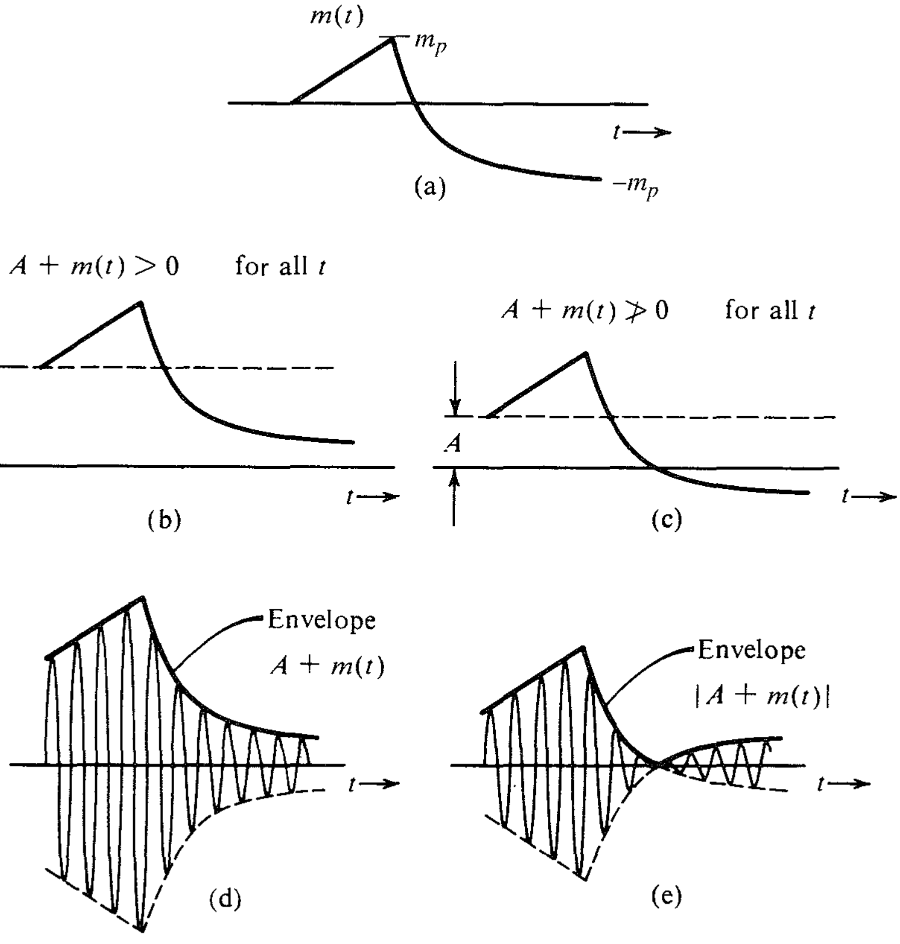


Figure 4.8 AM signal and its envelope.

---

[*] $E(t)$ must also be a slowly varying signal as compared to $\cos \omega_c t$.

Fig. 4.8d amounts to simple envelope detection. Thus, the condition for envelope detection of an AM signal is

$$A + m(t) \geq 0 \qquad \text{for all } t \tag{4.9a}$$

If $m(t) \geq 0$ for all t, then $A = 0$ also satisfies the condition (4.9a). In this case there is no need to add any carrier because the envelope of the DSB-SC signal $m(t)\cos\omega_c t$ is $m(t)$ and such a DSB-SC signal can be detected by envelope detection. In the following discussion we assume that $m(t) \not\geq 0$ for all t, that is, $m(t)$ does take on negative values over some range of t.

Let m_p be the peak amplitude (positive or negative) of $m(t)$ (see Fig. 4.8). This means that $m(t) \geq -m_p$. Hence, the condition (4.9a) is equivalent to*

$$A \geq m_p \tag{4.9b}$$

Thus, the minimum carrier amplitude required for the viability of envelope detection is m_p. This is quite clear from Fig. 4.8.

We define the modulation index μ as

$$\mu = \frac{m_p}{A} \tag{4.10a}$$

where A is the carrier amplitude. Note that m_p is a constant of the signal $m(t)$. Because $A \geq m_p$ and because there is no upper bound on A, it follows that

$$0 \leq \mu \leq 1 \tag{4.10b}$$

as the required condition for the viability of demodulation of AM by an envelope detector.

When $A < m_p$, Eq. (4.10a) shows that $\mu > 1$ (overmodulation). In this case, the option of envelope detection is no longer viable. We then need to use synchronous demodulation. Note that synchronous demodulation can be used for any value of μ (see Prob. 4.3-1). The envelope detector, which is considerably simpler and less expensive than the synchronous detector, can be used only for $\mu \leq 1$.

EXAMPLE 4.4 Sketch $\varphi_{\text{AM}}(t)$ for modulation indices of $\mu = 0.5$ and $\mu = 1$, when $m(t) = B\cos\omega_m t$. This case is referred to as **tone modulation** because the modulating signal is a pure sinusoid (or tone).

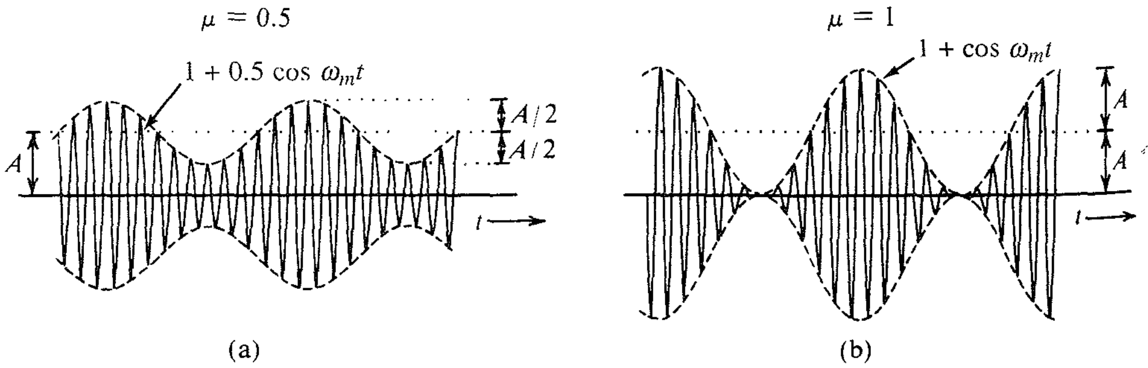


Figure 4.9 Tone-modulated AM. (a) $\mu = 0.5$. (b) $\mu = 1$.

---

* In case the negative and the positive peak amplitudes are not identical, m_p in condition (4.9b) is the absolute negative peak amplitude.

In this case, $m_p = B$ and the modulation index according to Eq. (4.10a) is

$$\mu = \frac{B}{A}$$

Hence, $B = \mu A$ and

$$m(t) = B\cos\omega_m t = \mu A\cos\omega_m t$$

Therefore,

$$\varphi_{\text{AM}}(t) = [A + m(t)]\cos\omega_c t = A[1 + \mu\cos\omega_m t]\cos\omega_c t \tag{4.11}$$

Figure 4.9 shows the modulated signals corresponding to $\mu = 0.5$ and $\mu = 1$, respectively.

Sideband and Carrier Power

The advantage of envelope detection in AM has its price. In AM, the carrier term does not carry any information, and hence, the carrier power is wasted,

$$\varphi_{\text{AM}}(t) = \underbrace{A\cos\omega_c t}_{\text{carrier}} + \underbrace{m(t)\cos\omega_c t}_{\text{sidebands}}$$

The carrier power P_c is the mean square value of $A\cos\omega_c t$, which is $A^2/2$. The sideband power P_s is the power of $m(t)\cos\omega_c t$, which is $0.5\,\widetilde{m^2(t)}$ [see Eq. (3.70)]. Hence,

$$P_c = \frac{A^2}{2} \quad \text{and} \quad P_s = \frac{1}{2}\widetilde{m^2(t)}$$

The sideband power is the useful power and the carrier power is the power wasted for convenience. The total power is the sum of the carrier (wasted) power and the sideband (useful) power. Hence, η, the power efficiency, is

$$\eta = \frac{\text{useful power}}{\text{total power}} = \frac{P_s}{P_c + P_s} = \frac{\widetilde{m^2(t)}}{A^2 + \widetilde{m^2(t)}}100\%$$

For the special case of tone modulation,

$$m(t) = \mu A\cos\omega_m t \quad \text{and} \quad \widetilde{m^2(t)} = \frac{(\mu A)^2}{2}$$

Hence

$$\eta = \frac{\mu^2}{2 + \mu^2}100\%$$

with the condition that $0 \le \mu \le 1$. It can be seen that η increases monotonically with μ, and η_{max} occurs at $\mu = 1$, for which

$$\eta_{\text{max}} = 33\%$$

Thus, for tone modulation, under best conditions ($\mu = 1$), only one-third of the transmitted power is used for carrying message. For practical signals, the efficiency is even worse—on the order of 25% or lower—compared to that of the DSB-SC case. The best condition implies

$\mu = 1$. Smaller values of μ degrade efficiency further. For this reason volume compression and peak limiting are commonly used in AM to ensure that full modulation ($\mu = 1$) is maintained most of the time.

EXAMPLE 4.5 Determine η and the percentage of the total power carried by the sidebands of the AM wave for tone modulation when **(a)** $\mu = 0.5$ and **(b)** $\mu = 0.3$.

For $\mu = 0.5$,

$$\eta = \frac{\mu^2}{2+\mu^2}100\% = \frac{(0.5)^2}{2+(0.5)^2}100\% = 11.11\%$$

Hence, only about 11% of the total power is in the sidebands. For $\mu = 0.3$,

$$\eta = \frac{(0.3)^2}{2+(0.3)^2}100\% = 4.3\%$$

Hence, only 4.3% of the total power is the useful power (power in sidebands).

Generation of AM Signals

AM signals can be generated by any DSB-SC modulators discussed in Sec. 4.2 if the modulating signal is $A + m(t)$ instead of just $m(t)$. But because there is no need to suppress the carrier in the output, the modulating circuits do not have to be balanced. This results in considerably simpler modulators for AM. Figure 4.10 shows a switching modulator, where the switching action is provided by a single diode (instead of a diode bridge as in Fig. 4.5). The input is $c \cos \omega_c t + m(t)$ with $c \gg m(t)$, so that the switching action of the diode is controlled by $c \cos \omega_c t$. The diode opens and shorts periodically with $\cos \omega_c t$, in effect multiplying the input signal $[c \cos \omega_c t + m(t)]$ by $w(t)$. The voltage across terminals bb' is

$$\begin{aligned} v_{bb'}(t) &= [c \cos \omega_c t + m(t)]w(t) \\ &= [c \cos \omega_c t + m(t)]\left[\frac{1}{2} + \frac{2}{\pi}\left(\cos \omega_c t - \frac{1}{3}\cos 3\omega_c t + \frac{1}{5}\cos 5\omega_c t - \cdots\right)\right] \\ &= \underbrace{\frac{c}{2}\cos \omega_c t + \frac{2}{\pi}m(t)\cos \omega_c t}_{\text{AM}} + \underbrace{\text{other terms}}_{\substack{\text{suppressed by} \\ \text{bandpass filter}}} \end{aligned}$$

The bandpass filter tuned to ω_c suppresses all the other terms, yielding the desired AM signal at the output.

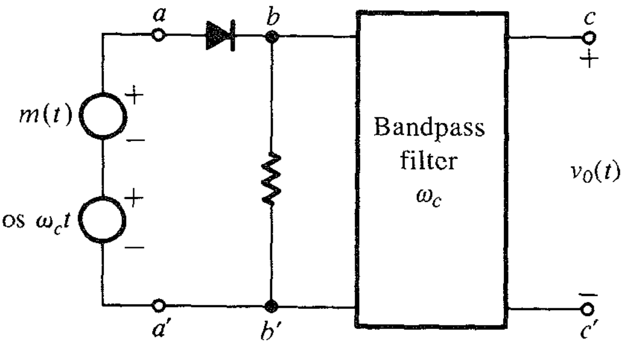


Figure 4.10 AM generator.

Demodulation of AM Signals

The AM signal can be demodulated coherently by a locally generated carrier (see Prob. 4.3-1). However, coherent, or synchronous, demodulation of AM* will defeat the very purpose of AM and, hence, is rarely used in practice. We shall consider here two noncoherent methods of AM demodulation: (1) rectifier detection, and (2) envelope detection.

Rectifier Detector: If an AM signal is applied to a diode and a resistor circuit (Fig. 4.11), the negative part of the AM wave will be suppressed. The output across the resistor is a half-wave rectified version of the AM signal. In essence, the AM signal is multiplied by $w(t)$. Hence, the rectified output v_R is

$$\begin{aligned} v_R &= \{[A + m(t)] \cos \omega_c t\}\, w(t) \\ &= [A + m(t)] \cos \omega_c t \left[\frac{1}{2} + \frac{2}{\pi}\left(\cos \omega_c t - \frac{1}{3} \cos 3\omega_c t + \frac{1}{5} \cos 5\omega_c t - \cdots\right)\right] \\ &= \frac{1}{\pi}[A + m(t)] + \text{other terms of higher frequencies} \end{aligned}$$

When v_R is applied to a low-pass filter of cutoff B Hz, the output is $[A + m(t)]/\pi$, and all the other terms in v_R of frequencies higher than B Hz are suppressed. The dc term A/π may be blocked by a capacitor (Fig. 4.11) to give the desired output $m(t)/\pi$. The output can be doubled by using a full-wave rectifier.

It is interesting to note that rectifier detection is in effect synchronous detection performed without using a local carrier. The high carrier content in AM ensures that its zero crossings are periodic and the information about frequency and phase of the carrier at the transmitter is built in to the AM signal itself.

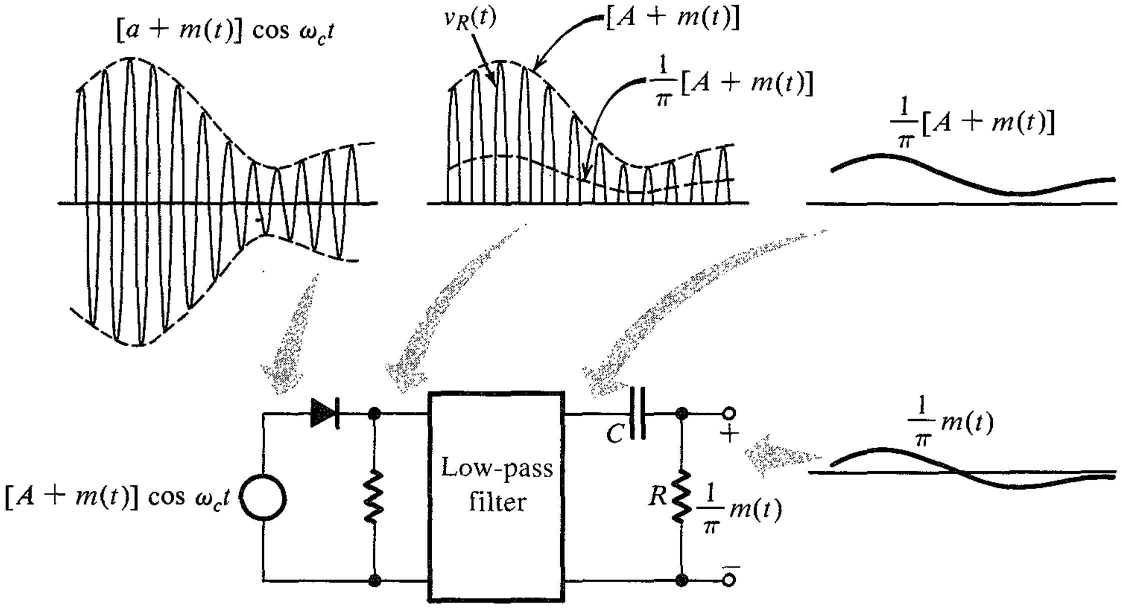

$[a + m(t)] \cos \omega_c t$
$v_R(t)$
$[A + m(t)]$
$\frac{1}{\pi}[A + m(t)]$
$\frac{1}{\pi}[A + m(t)]$
$\frac{1}{\pi}m(t)$
$[A + m(t)] \cos \omega_c t$
Low-pass filter
C
+
R
$\frac{1}{\pi}m(t)$
−

Figure 4.11 Rectifier detector for AM.

---

* By AM, we mean the case $\mu \leq 1$.

Envelope Detector: In an envelope detector, the output of the detector follows the envelope of the modulated signal. The circuit shown in Fig. 4.12a functions as an envelope detector. On the positive cycle of the input signal, the diode conducts and the capacitor C charges up to the peak voltage of the input signal. As the input signal falls below this peak value, the diode is cut off, because the capacitor voltage (which is very nearly the peak voltage) is greater than the input signal voltage, thus causing the diode to open. The capacitor now discharges through the resistor R at a slow rate (with a time constant RC). During the next positive cycle, the same drama repeats. when the input signal becomes greater than the capacitor voltage, the diode conducts again. The capacitor again charges to the peak value of this (new) cycle. The capacitor discharges slowly during the cutoff period, thus changing the capacitor voltage very slightly.

During each positive cycle, the capacitor charges up to the peak voltage of the input signal and then decays slowly until the next positive cycle as shown in Fig. 4.12b. The output voltage $v_C(t)$, thus, closely follows the envelope of the input. Capacitor discharge between positive peaks causes a ripple signal of frequency ω_c in the output. This ripple can be reduced by increasing the time constant RC so that the capacitor discharges very little between the positive peaks ($RC \gg 1/\omega_c$). Making RC too large, however, would make it impossible for the capacitor voltage to follow the envelope (see Fig. 4.12b). Thus, RC should be large

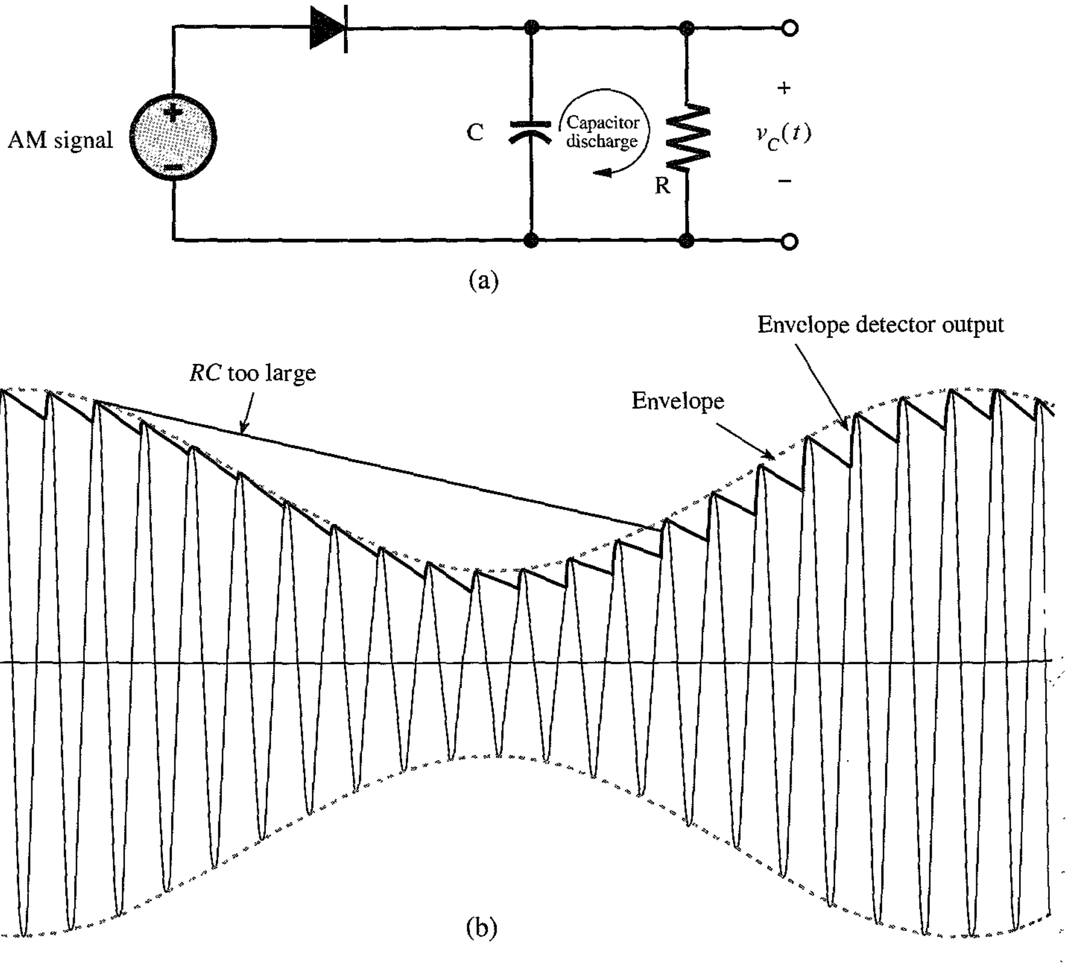


Figure 4.12 Envelope detector for AM.

compared to $1/\omega_c$ but should be small compared to $1/2\pi B$, where B is the highest frequency in $m(t)$ (see Example 4.6). This, incidentally, also requires that $\omega_c \gg 2\pi B$, a condition that is necessary for a well-defined envelope.

The envelope-detector output is $v_C(t) = A + m(t)$ with a ripple of frequency ω_c. The dc term A can be blocked out by a capacitor or a simple RC high-pass filter. The ripple may be reduced further by another (low-pass) RC filter.

Both the rectifier detector and the envelope detector consist of a half-wave rectifier followed by a low-pass filter. Superficially, these detectors may appear equivalent, but they are distinct and operate on very different principles. The rectifier detector is basically a synchronous demodulator. The envelope detection, on the other-hand, is a nonlinear operation. Observe that the low-pass filter in the rectifier detector is designed to separate $m(t)$ from terms such as $m(t)\cos n\omega_c t$; it does not depend on the value of μ. On the other hand, we show in Example 4.6 that the time constant RC of the low-pass filter for the envelope detector does depend on the value of μ.

EXAMPLE 4.6 For tone modulation (Example 4.4), determine the upper limit of RC to ensure that the capacitor voltage follows the envelope.

Figure 4.13 shows the envelope and the voltage across the capacitor. The capacitor discharges from the peak value E starting at some arbitrary instant $t = 0$. The voltage v_C across the capacitor is given by

$$v_C = Ee^{-t/RC}$$

Because the time constant is much larger than the interval between the two successive cycles of the carrier ($RC \gg 1/\omega_c$), the capacitor voltage v_C discharges exponentially for a short time compared to its time constant. Hence, the exponential can be approximated by a straight line obtained from the first two terms in Taylor's series for $Ee^{-t/RC}$,

$$v_C \simeq E\left(1 - \frac{t}{RC}\right)$$

The slope of the discharge is $-E/RC$. In order for the capacitor to follow the envelope $E(t)$, the magnitude of the slope of the RC discharge must be greater than the magnitude of the slope of the envelope $E(t)$. Hence,

$$\left|\frac{dv_C}{dt}\right| = \frac{E}{RC} \geq \left|\frac{dE}{dt}\right| \tag{4.12}$$

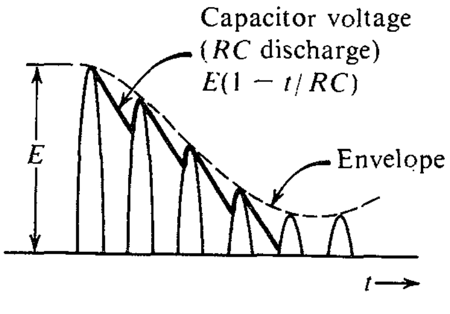


Figure 4.13 Capacitor discharge in an envelope detector.

But the envelope $E(t)$ of a tone-modulated carrier is [Eq. (4.11)]

$$E(t) = A[1 + \mu \cos \omega_m t]$$
$$\frac{dE}{dt} = -\mu A \omega_m \sin \omega_m t$$

Hence, Eq. (4.12) becomes

$$\frac{A(1 + \mu \cos \omega_m t)}{RC} \geq \mu A \omega_m \sin \omega_m t \qquad \text{for all } t$$

or

$$RC \leq \frac{1 + \mu \cos \omega_m t}{\mu \omega_m \sin \omega_m t} \qquad \text{for all } t$$

The worst possible case occurs when the right-hand side is the minimum. This is found (as usual, by taking the derivative and setting it to zero) to be when $\cos \omega_m t = -\mu$. For this case, the right-hand side is $\sqrt{(1 - \mu^2)}/\mu\omega_m$. Hence,

$$RC \leq \frac{1}{\omega_m}\left(\frac{\sqrt{1 - \mu^2}}{\mu}\right)$$

4.4 QUADRATURE AMPLITUDE MODULATION (QAM)

The DSB signals occupy twice the bandwidth required for the baseband. This disadvantage can be overcome by transmitting two DSB signals using carriers of the same frequency but in phase quadrature, as shown in Fig. 4.14. In this figure, the boxes labeled $-\pi/2$ are phase shifters, which delay the phase of an input sinusoid by $-\pi/2$ rad. If the two baseband signals to be transmitted are $m_1(t)$ and $m_2(t)$, the corresponding QAM signal $\varphi_{\text{QAM}}(t)$, the sum of the two DSB-modulated signals, is

$$\varphi_{\text{QAM}}(t) = m_1(t) \cos \omega_c t + m_2(t) \sin \omega_c t$$

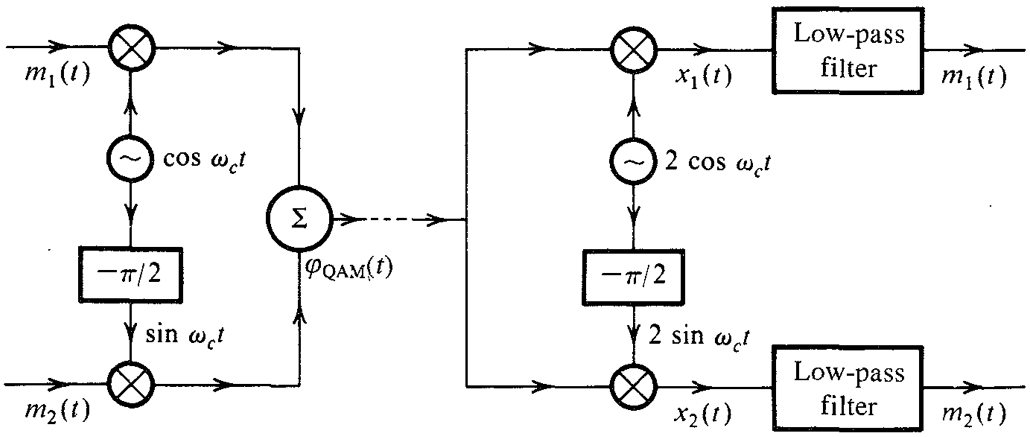


Figure 4.14 Quadrature amplitude multiplexing.

Both modulated signals occupy the same band. Yet two baseband signals can be separated at the receiver by synchronous detection using two local carriers in phase quadrature, as shown in Fig. 4.14. This can be shown by considering the multiplier output $x_1(t)$ of the upper arm of the receiver (Fig. 4.14):

$$\begin{aligned} x_1(t) &= 2\varphi_{\text{QAM}}(t)\cos\omega_c t = 2[m_1(t)\cos\omega_c t + m_2(t)\sin\omega_c t]\cos\omega_c t \\ &= m_1(t) + m_1(t)\cos 2\omega_c t + m_2(t)\sin 2\omega_c t \end{aligned}$$

The last two terms are suppressed by the low-pass filter, yielding the desired output $m_1(t)$. Similarly, the output of the lower receiver branch can be shown to be $m_2(t)$. This scheme is known as **quadrature amplitude modulation (QAM)** or **quadrature multiplexing**. Thus, two baseband signals, each of bandwidth B Hz, can be transmitted simultaneously over a bandwidth $2B$ by using DSB transmission and quadrature multiplexing. The upper channel is also known as the **in-phase** (I) channel and the lower channel is the **quadrature** (Q) channel.

QAM is somewhat of an exacting scheme. A slight error in the phase or the frequency of the carrier at the demodulator in QAM will not only result in loss and distortion of signals, but will also lead to interference between the two channels. To show this let the carrier at the demodulator be $2\cos(\omega_c t + \theta)$. In this case,

$$\begin{aligned} x_1(t) &= 2[m_1(t)\cos\omega_c t + m_2(t)\sin\omega_c t]\cos(\omega_c t + \theta) \\ &= m_1(t)\cos\theta + m_1(t)\cos(2\omega_c t + \theta) - m_2(t)\sin\theta + m_2(t)\sin(2\omega_c t + \theta) \end{aligned}$$

The low-pass filter suppresses the two signals with frequency $2\omega_c$, resulting in the output $m_1(t)\cos\theta - m_2(t)\sin\theta$. Thus, in addition to the desired signal $m_1(t)$, we also receive signal $m_2(t)$ in the upper branch. Similar argument shows that in addition to the desired signal $m_2(t)$, we receive signal $m_1(t)$ in the lower branch. This cochannel* interference is undesirable. Similar difficulties arise when the local frequency is in error (see Prob. 4.4-1). In addition, unequal attenuation of the USB and the LSB during transmission also leads to crosstalk or cochannel interference.

Quadrature multiplexing is used in color television to multiplex the so-called chrominance signals, which carry the information about colors. There the synchronization is achieved by periodic insertion of a short burst of carrier signal (called **color burst** in the transmitted signal, as explained in Sec. 4.9).

4.5 AMPLITUDE MODULATION: SINGLE SIDEBAND (SSB)

The DSB spectrum has two sidebands: the upper sideband (USB) and the lower sideband (LSB), both containing the complete information of the baseband signal (Fig. 4.15). A scheme in which only one sideband is transmitted is known as **single-sideband (SSB) transmission**, which requires only one-half the bandwidth of the DSB signal.

An SSB signal can be coherently (synchronously) demodulated. For example, multiplication of a USB signal (Fig. 4.15c) by $\cos\omega_c t$ shifts its spectrum to the left and right by ω_c, yielding the spectrum in Fig. 4.15e. Low-pass filtering of this signal yields the desired baseband signal. The case is similar with LSB signals. Hence, demodulation of SSB signals

---

* Cochannel refers to channels having the same carrier frequency.

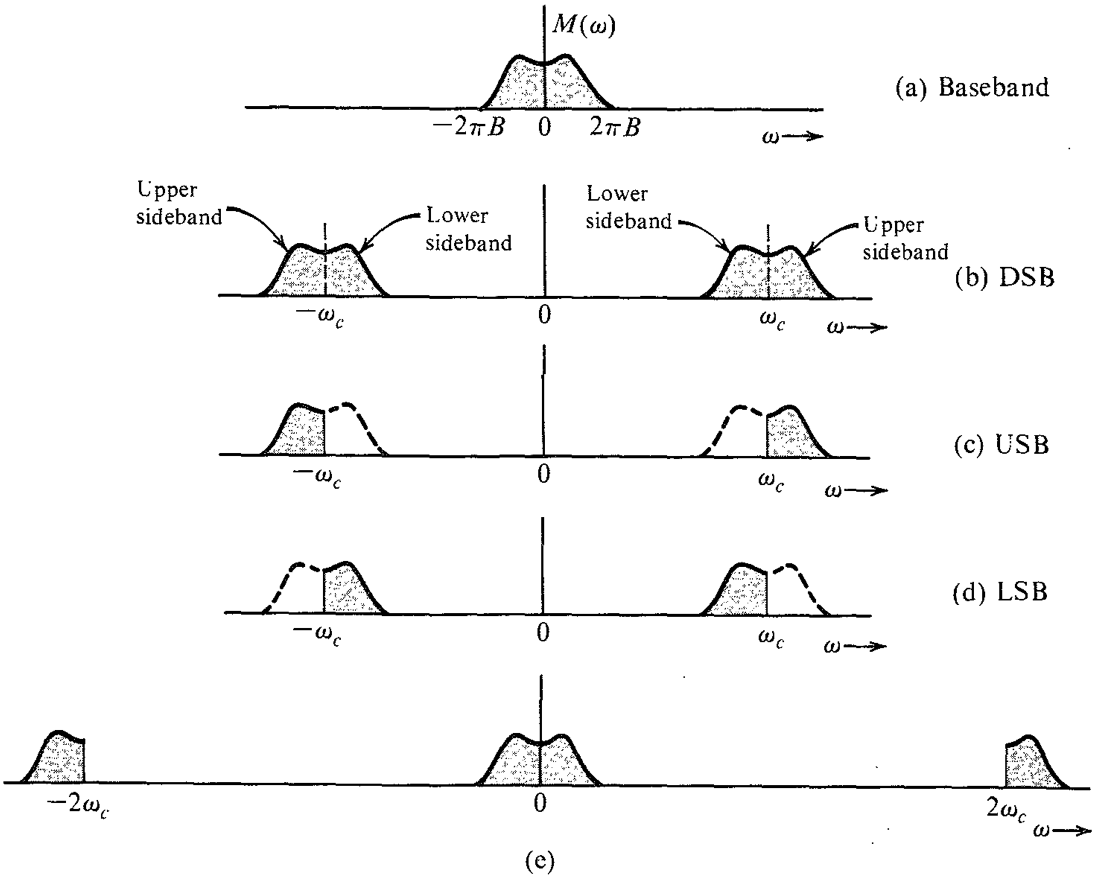

M(ω)
(a) Baseband
−2πB
0
2πB
ω→
Upper sideband
Lower sideband
Lower sideband
Upper sideband
(b) DSB
−ω_c
0
ω_c
ω→
(c) USB
−ω_c
0
ω_c
ω→
(d) LSB
−ω_c
0
ω_c
ω→
−2ω_c
0
2ω_c
ω→
(e)

Figure 4.15 SSB spectra.

is identical to that of DSB-SC signals. Note that we are talking of SSB signals without an additional carrier, and, hence, they are suppressed carrier signals (SSB-SC).

Time-Domain Representation of SSB Signals

Because the building blocks of an SSB signal are the sidebands, we shall first obtain a time-domain expression for each sideband. Figure 4.16a shows the spectrum $M(\omega)$. Figure 4.16b shows the USB $M_+(\omega)$ and Fig. 4.16c shows the LSB $M_-(\omega)$. From Fig. 4.16b and c, we observe that $M_+(\omega) = M(\omega)u(\omega)$ and $M_-(\omega) = M(\omega)u(-\omega)$. Let $m_+(t)$ and $m_-(t)$ be the inverse Fourier transforms of $M_+(\omega)$ and $M_-(\omega)$, respectively.* Because the amplitude spectra $|M_+(\omega)|$ and $|M_-(\omega)|$ are not even functions of ω, the signals $m_+(t)$ and $m_-(t)$ cannot be real; they are complex. Moreover, $M_+(\omega)$ and $M_-(\omega)$ are the two halves of $M(\omega)$. Hence, from Eqs. (3.10), it follows that $M_+(-\omega)$ and $M_-(\omega)$ are conjugates. Consequently, $m_+(t)$ and $m_-(t)$ are conjugates (see Prob. 3.1-3). Also, because $m_+(t) + m_-(t) = m(t)$, we can express

$$m_+(t) = \frac{1}{2}[m(t) + jm_h(t)] \tag{4.13a}$$

and

$$m_-(t) = \frac{1}{2}[m(t) - jm_h(t)] \tag{4.13b}$$

---

* In the literature, $2m_+(t)$ is also known as the preenvelope of $m(t)$.

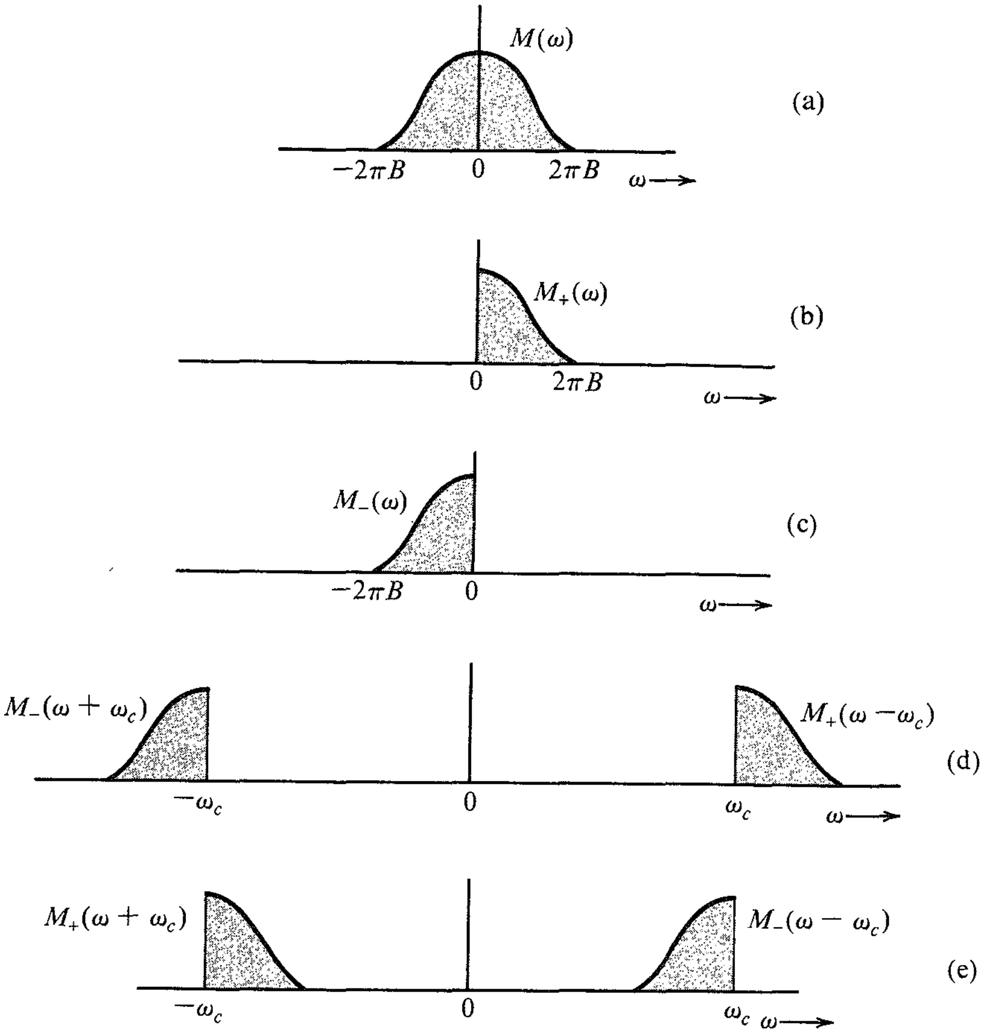

$M(\omega)$
(a)
$-2\pi B$
0
$2\pi B$
$\omega \longrightarrow$
$M_+(\omega)$
(b)
0
$2\pi B$
$\omega \longrightarrow$
$M_-(\omega)$
(c)
$-2\pi B$
0
$\omega \longrightarrow$
$M_-(\omega + \omega_c)$
$M_+(\omega - \omega_c)$
(d)
$-\omega_c$
0
ω_c
$\omega \longrightarrow$
$M_+(\omega + \omega_c)$
$M_-(\omega - \omega_c)$
(e)
$-\omega_c$
0
ω_c
$\omega \longrightarrow$

Figure 4.16 Expressing SSB spectra in terms of $M_+(\omega)$ and $M_-(\omega)$.

where $m_h(t)$ is unknown. To determine $m_h(t)$, we note that

$$\begin{aligned} M_+(\omega) &= M(\omega)u(\omega) \\ &= \tfrac{1}{2}M(\omega)[1 + \operatorname{sgn}(\omega)] \\ &= \tfrac{1}{2}M(\omega) + \tfrac{1}{2}M(\omega)\operatorname{sgn}(\omega) \end{aligned} \tag{4.14a}$$

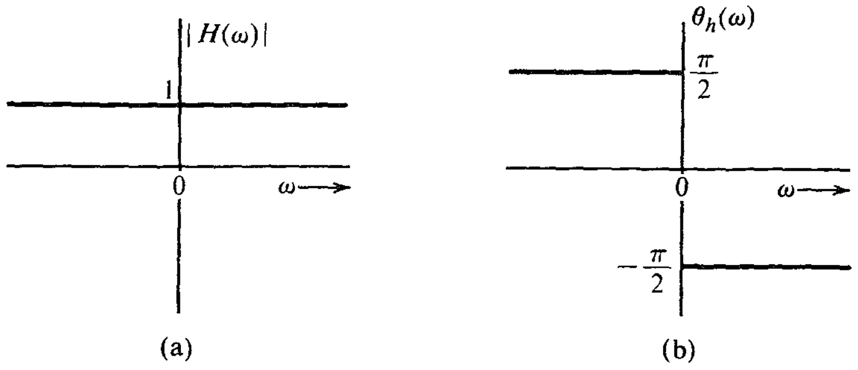
From Eqs. (4.13a) and (4.14a), it follows that $jm_h(t) \Leftrightarrow M(\omega)\ \operatorname{sgn}(\omega)$. Hence,

$$M_h(\omega) = -jM(\omega)\operatorname{sgn}(\omega) \tag{4.14b}$$

Application of the duality property to pair 12 of Table 3.1 yields $1/\pi t \Longleftrightarrow -j\operatorname{sgn}(\omega)$. Applying this result and the time convolution property to Eq. (4.14b) yields $m_h(t) = m(t) * 1/\pi t$, that is,

$$m_h(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{m(\alpha)}{t-\alpha}\,d\alpha \tag{4.15}$$

The right-hand side of Eq. (4.15) defines the **Hilbert transform** of $m(t)$. Thus, the signal $m_h(t)$



(a) (b)

Figure 4.17 Transfer function of an ideal $\pi/2$ phase shifter (Hilbert transformer).

is the Hilbert transform of $m(t)$. From Eq. (4.14b), it follows that if $m(t)$ is passed through a transfer function $H(\omega) = -j\ \text{sgn}\ (\omega)$, then the output is $m_h(t)$, the Hilbert transform of $m(t)$. Because

$$H(\omega) = -j\ \text{sgn}\ (\omega) \tag{4.16a}$$

$$= \begin{cases} -j = 1e^{-j\pi/2} & \omega > 0 \\ j = 1e^{j\pi/2} & \omega < 0 \end{cases} \tag{4.16b}$$

it follows that $|H(\omega)| = 1$ and that $\theta_h(\omega) = -\pi/2$ for $\omega > 0$ and $\pi/2$ for $\omega < 0$, as shown in Fig. 4.17. Thus, if we delay the phase of every component of $m(t)$ by $\pi/2$ (without changing its amplitude), the resulting signal is $m_h(t)$, the Hilbert transform of $m(t)$. Therefore, a Hilbert transformer is an ideal phase shifter that shifts the phase of every spectral component by $-\pi/2$.

We can now express the SSB signal in terms of $m(t)$ and $m_h(t)$. From Fig. 4.16d it is clear that the USB spectrum $\Phi_{\text{USB}}(\omega)$ can be expressed as

$$\Phi_{\text{USB}}(\omega) = M_+(\omega - \omega_c) + M_-(\omega + \omega_c)$$

The inverse transform of this equation yields

$$w\varphi_{\text{USB}}(t) = m_+(t)e^{j\omega_c t} + m_-(t)e^{-j\omega_c t}$$

Substituting Eqs. (4.13) in the preceding equation yields

$$\varphi_{\text{USB}}(t) = m(t)\cos\omega_c t - m_h(t)\sin\omega_c t \tag{4.17a}$$

Using a similar argument, we can show that

$$\varphi_{\text{LSB}}(t) = m(t)\cos\omega_c t + m_h(t)\sin\omega_c t \tag{4.17b}$$

Hence, a general SSB signal $\varphi_{\text{SSB}}(t)$ can be expressed as

$$\varphi_{\text{SSB}}(t) = m(t)\cos\omega_c t \mp m_h(t)\sin\omega_c t \tag{4.17c}$$

where the minus sign applies to USB and the plus sign applies to LSB.

EXAMPLE 4.7 **Tone Modulation: SSB**

Find $\varphi_{\text{SSB}}(t)$ for a simple case of a tone modulation, that is, when the modulating signal is a sinusoid $m(t) = \cos\omega_m t$.

Recall that the Hilbert transform delays the phase of each spectral component by $\pi/2$. In the present case, there is only one spectral component of frequency ω_m. Delaying the phase of $m(t)$ by $\pi/2$ yields

$$m_h(t) = \cos\left(\omega_m t - \frac{\pi}{2}\right) = \sin \omega_m t$$

Hence, from Eq. (4.17c),

$$\begin{aligned}\varphi_{\text{SSB}}(t) &= \cos \omega_m t \cos \omega_c t \mp \sin \omega_m t \sin \omega_c t \\ &= \cos (\omega_c \pm \omega_m)t\end{aligned}$$

Thus,

$$\varphi_{\text{USB}}(t) = \cos (\omega_c + \omega_m)t \qquad \varphi_{\text{LSB}}(t) = \cos (\omega_c - \omega_m)t$$

To verify these results, consider the spectrum of $m(t)$ (Fig. 4.18a) and its DSB-SC (Fig. 4.18b), USB (Fig. 4.18c), and LSB (Fig. 4.18d) spectra. It is evident that the spectra in Fig. 4.18c and d do indeed correspond to the $\varphi_{\text{USB}}(t)$ and $\varphi_{\text{LSB}}(t)$ derived earlier.

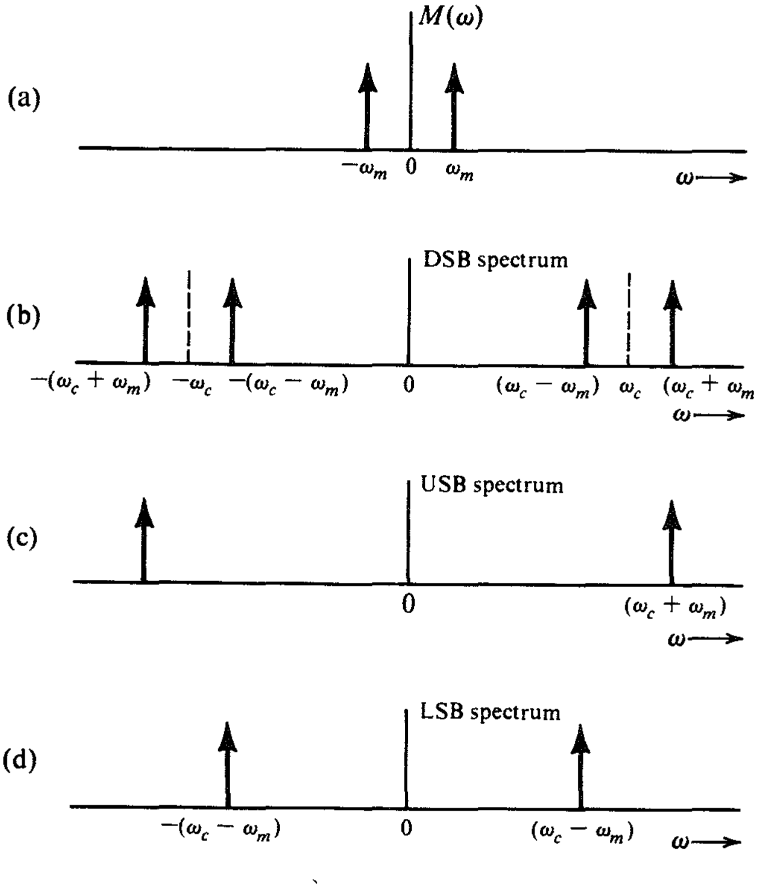

(a)
$M(\omega)$
$-\omega_m$
0
ω_m
$\omega \longrightarrow$
(b)
DSB spectrum
$-(\omega_c + \omega_m)$
$-\omega_c$
$-(\omega_c - \omega_m)$
0
$(\omega_c - \omega_m)$
ω_c
$(\omega_c + \omega_m)$
$\omega \longrightarrow$
(c)
USB spectrum
0
$(\omega_c + \omega_m)$
$\omega \longrightarrow$
(d)
LSB spectrum
$-(\omega_c - \omega_m)$
0
$(\omega_c - \omega_m)$
$\omega \longrightarrow$

Figure 4.18 SSB spectra for tone modulation.

Generation of SSB Signals[2]

Two methods are commonly used to generate SSB signals. The first method uses sharp cutoff filters to eliminate the undesired sideband, and the second method uses phase-shifting networks to achieve the same goal. Yet another method, known as Weaver's method,[3] can also be used to generate SSB signals, provided the baseband signal spectrum has little power near the origin.

Selective-Filtering Method: This is the most commonly used method of generating SSB signals. In this method, a DSB-SC signal is passed through a sharp cutoff filter to eliminate the undesired sideband.

To obtain the USB, the filter should pass all components above ω_c unattenuated and completely suppress all components below ω_c. Such an operation requires an ideal filter, which is unrealizable. It can, however, be realized closely if there is some separation between the passband and the stopband. Fortunately, the voice signal provides this condition, because its spectrum shows little power content at the origin (Fig. 4.19a). In addition, articulation tests have shown that for speech signals, frequency components below 300 Hz are not important. In other words, we may suppress all speech components below 300 Hz without affecting the intelligibility appreciably.* Thus, filtering of the unwanted sideband becomes relatively easy for speech signals because we have a 600-Hz transition region around the cutoff frequency ω_c. To minimize adjacent channel interference, the undesired sideband should be attenuated at least 40 dB.†

Phase-Shift Method: Equation (4.17) is the basis for this method. Figure 4.20 shows the implementation of Eq. (4.17). The box marked "$-\pi/2$" is a $\pi/2$ phase shifter, which delays the phase of every spectral component by $\pi/2$. Hence, it is a Hilbert transformer. Note that

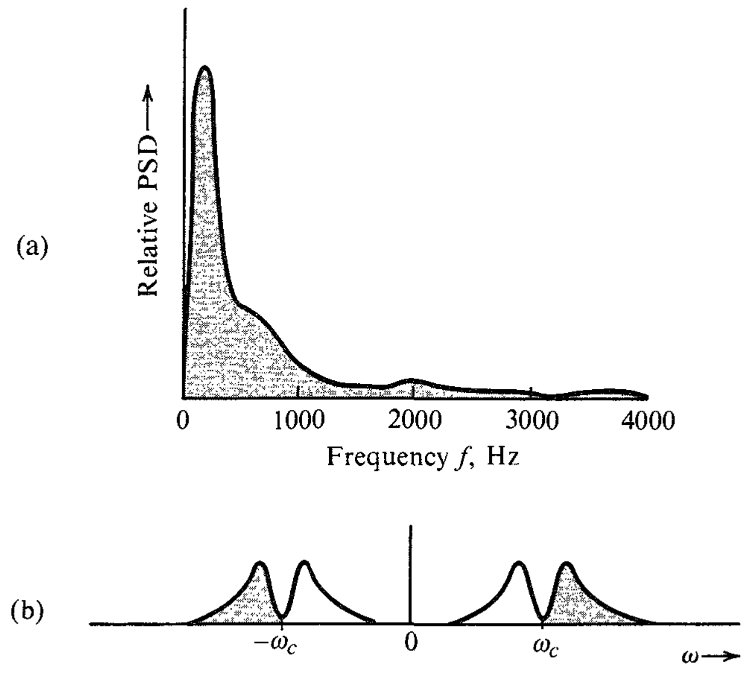


Figure 4.19 Relative power spectrum of speech signal and the corresponding USB spectrum.

---

* Similarly, suppression of speech-signal components above 3500 Hz causes no appreciable change in intelligibility.

† For very high carrier frequencies, the ratio of the gap band (600 Hz) to the carrier frequency may be too small, and, thus, a transition of 40 dB in amplitude over 600 Hz may pose a problem. In such a case, the modulation is carried out using a smaller carrier frequency (ω_{c_1}) first. The resulting SSB signal effectively widens the gap to $2\omega_{c_1}$ (see Fig. 4.19c). Now, treating this signal as the new baseband signal, it is possible to SSB-modulate the high-frequency carrier.

an ideal phase shifter is also unrealizable. We can, at most, approximate it over a finite band. However, it is possible to realize a filter with two outputs such that both outputs have the same (constant) amplitude spectrum, but their phase spectra differ by $\pi/2$ rad over a given band of frequencies.*

In terms of bandwidth requirement, SSB is similar to QAM but less exacting in terms of the carrier frequency and phase or the requirement of a distortionless transmission medium. However, SSB is difficult to generate if the baseband signal has no dc null in its spectrum. It is easy to build a circuit to shift the phase of a single frequency component by $\pi/2$ rad. But a device to achieve a $\pi/2$ phase shift of all the spectral components over a band of frequencies is unrealizable. We can, at best, approximate it over a finite band.

Demodulation of SSB-SC Signals

It was shown earlier that SSB-SC signals can be coherently demodulated. We can readily verify this in another way:

$$\varphi_{\text{SSB}}(t) = m(t)\cos\omega_c t \mp m_h(t)\sin\omega_c t$$

Hence,

$$\begin{aligned}\varphi_{\text{SSB}}(t)\cos\omega_c t &= \tfrac{1}{2}m(t)[1+\cos 2\omega_c t] \mp \tfrac{1}{2}m_h(t)\sin 2\omega_c t\\ &= \tfrac{1}{2}m(t) + \tfrac{1}{2}[m(t)\cos 2\omega_c t \mp m_h(t)\sin 2\omega_c t]\end{aligned}$$

Thus, the product $\varphi_{\text{SSB}}(t)\cos\omega_c t$ yields the baseband signal and another SSB signal with a carrier $2\omega_c$. The spectrum in Fig. 4.15e shows precisely this result. A low-pass filter will suppress the unwanted SSB terms, giving the desired baseband signal $m(t)/2$. Hence, the demodulator is identical to the synchronous demodulator used for DSB-SC. Thus, any one of the synchronous DSB-SC demodulators discussed in Sec. 4.2 can be used to demodulate an SSB-SC signal.

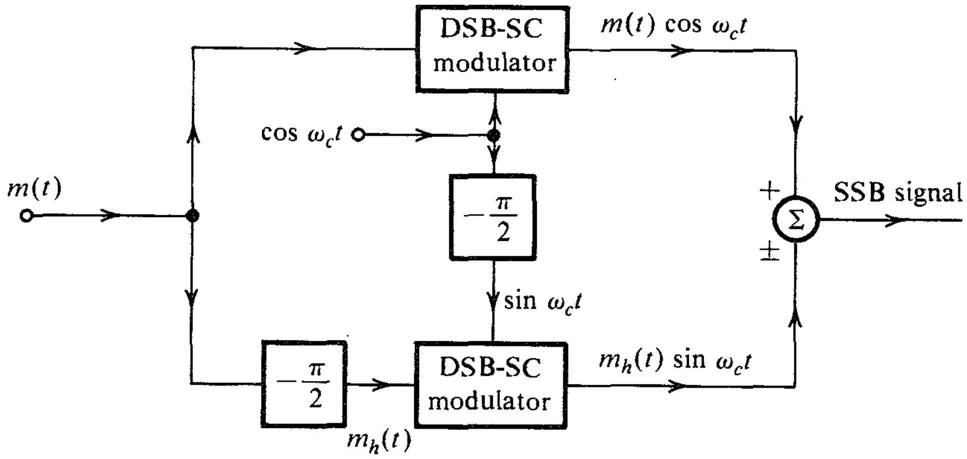


Figure 4.20 SSB generation by phase-shift method.

---

* In this case, the phase spectrum of one output may be $\phi(\omega)$ while that of the other is $\phi(\omega) - \pi/2$. The term $\phi(\omega)$ is an unwanted phase distortion. However, as seen earlier, human ear is not sensitive to this kind of distortion.

Envelope Detection of SSB Signals with a Carrier (SSB+C): We now consider SSB signals with an additional carrier (SSB+C). Such a signal can be expressed as

$$\varphi_{\text{SSB+C}} = A\cos\omega_c t + [m(t)\cos\omega_c t + m_h(t)\sin\omega_c t]$$

Although $m(t)$ can be recovered by synchronous detection [multiplying $\varphi_{\text{SSB+C}}$ by $\cos\omega_c t$] if A, the carrier amplitude, is large enough, $m(t)$ can also be recovered from $\varphi_{\text{SSB+C}}$ by envelope or rectifier detection. This can be shown by rewriting $\varphi_{\text{SSB+C}}$ as

$$\begin{aligned}\varphi_{\text{SSB+C}} &= [A + m(t)]\cos\omega_c t + m_h(t)\sin\omega_c t\\ &= E(t)\cos(\omega_c t + \theta)\end{aligned}$$

where $E(t)$, the envelope of $\varphi_{\text{SSB+C}}$, is given by [see Eq. (3.39)]

$$\begin{aligned}E(t) &= \{[A + m(t)]^2 + m_h^2(t)\}^{1/2}\\ &= A\left[1 + \frac{2m(t)}{A} + \frac{m^2(t)}{A^2} + \frac{m_h^2(t)}{A^2}\right]^{1/2}\end{aligned}$$

If $A \gg |m(t)|$, then in general* $A \gg |m_h(t)|$, and the terms $m^2(t)/A^2$ and $m_h^2(t)/A^2$ can be ignored. Thus,

$$E(t) \simeq A\left[1 + \frac{2m(t)}{A}\right]^{1/2}$$

Using binomial expansion and discarding higher order terms [because $m(t)/A \ll 1$], we get

$$\begin{aligned}E(t) &\simeq A\left[1 + \frac{m(t)}{A}\right]\\ &= A + m(t)\end{aligned}$$

It is evident that for a large carrier, the SSB + C can be demodulated by an envelope detector.

In AM, envelope detection requires the condition $A \geq |m(t)|$, whereas for SSB+C, the condition is $A \gg |m(t)|$. Hence, in SSB case, the required carrier amplitude is much larger than that in AM, and, consequently, the efficiency of SSB+C is pathetically low.

Telephone-Channel Multiplexing

Until recently, almost all long-haul telephone channels were multiplexed by FDM using SSB signals. This multiplexing technique, standardized by the CCITT, provides considerable flexibility in branching, dropping off, or inserting blocks of channels at points en route.[4] A basic **group** consists of 12 frequency-division multiplexed SSB voice channels, each of bandwidth 4 kHz (first-level multiplexing). A basic group uses LSB spectra and occupies a band of 60 to 108 kHz. An alternate group configuration of 12 USB voice signals, occupying a band of 148 to 196 kHz, is also used.

A basic **supergroup** of 60 channels is formed by multiplexing five basic groups, and it occupies a band of 312 to 552 kHz. An alternate supergroup configuration using USB spectra occupies a band of 60 to 300 kHz.

---

* This may not be true for all t, but it is true for most t.

A basic **mastergroup** of 600 channels is formed by multiplexing 10 supergroups.* There are two standard mastergroup configurations: the L600 and the U600.

Modern broad-band transmission systems can transmit even larger groupings than mastergroups. For the L3 carrier and TH microwave, three mastergroups and one supergroup comprising 1860 message channels are combined. The L4 system utilizes six U600 mastergroups multiplexed to form 3600 channels. The multiplexed signal is fed into the baseband input of a microwave radio channel or directly into a coaxial transmission system.

4.6 AMPLITUDE MODULATION: VESTIGIAL SIDEBAND (VSB)

As seen earlier, the generation of SSB signals is rather difficult. The selective-filtering method demands dc null in the modulating signal spectrum. A phase shifter required in the phase-shift method is unrealizable, or realizable only approximately. The generation of DSB signals is much simpler, but requires twice the signal bandwidth. A **vestigial-sideband (VSB)**, also called asymmetric sideband system is a compromise between DSB and SSB. It inherits the advantages of DSB and SSB but avoids their disadvantages at a small cost. VSB signals are relatively easy to generate, and, at the same time, their bandwidth is only (typically 25%) greater than that of SSB signals.

In VSB, instead of rejecting one sideband completely (as in SSB), a gradual cutoff of one sideband, as shown in Fig. 4.21d, is accepted. The baseband signal can be recovered exactly by a synchronous detector in conjunction with an appropriate equalizer filter $H_o(\omega)$ at the receiver output (Fig. 4.22). If a large carrier is transmitted along with the VSB signal, the baseband signal can be recovered by an envelope (or a rectifier) detector.

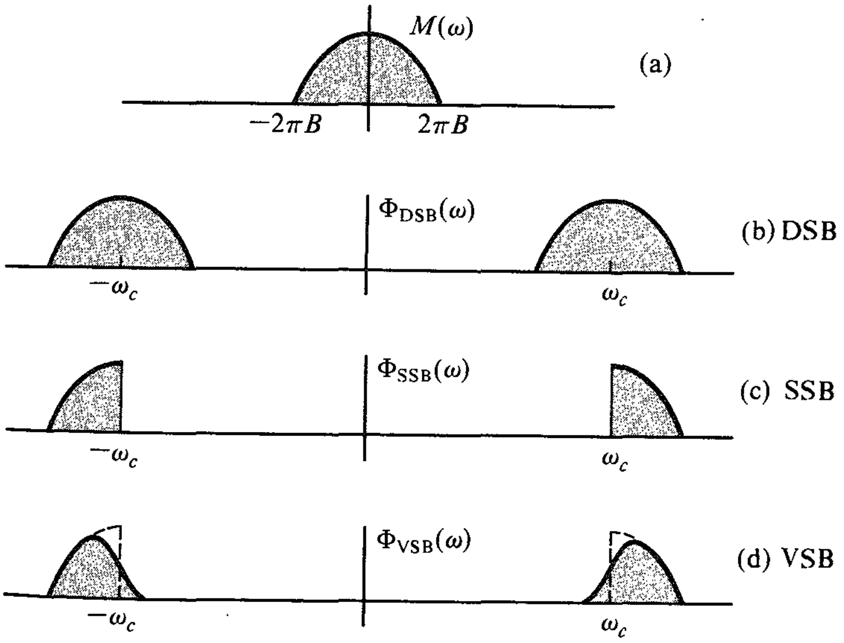


Figure 4.21 Spectra of the modulating signal and corresponding DSB, SSB, and VSB signals.

---

* This is true for the North American hierarchy. In the CCITT hierarchy, a basic mastergroup is formed by multiplexing five supergroups (300 voice channels).

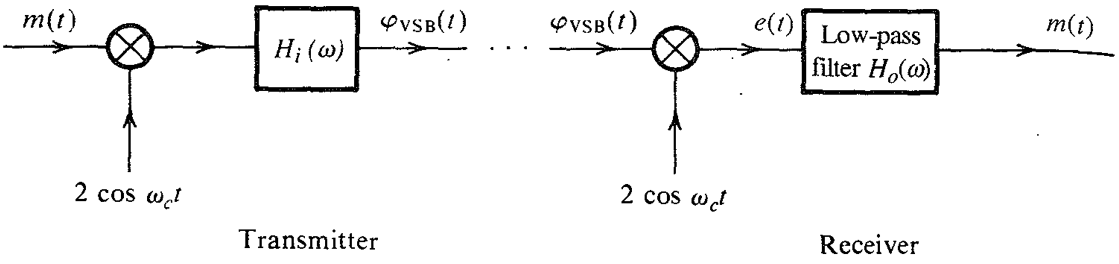


Figure 4.22 VSB modulator and demodulator.

If the vestigial shaping filter that produces VSB from DSB is $H_i(\omega)$ (Fig. 4.22), then the resulting VSB signal spectrum is

$$\Phi_{\text{VSB}}(\omega) = [M(\omega+\omega_c) + M(\omega-\omega_c)]H_i(\omega) \tag{4.18}$$

This VSB shaping filter $H_i(\omega)$ allows the transmission of one sideband, but suppresses the other sideband, not completely, but gradually. This makes it easy to realize such a filter, but the transmission bandwidth is now somewhat higher than that of the SSB (where the other sideband is suppressed completely). The bandwidth of the VSB signal is typically 25 to 33% higher than that of the SSB signals.

We require that $m(t)$ be recoverable from $\varphi_{\text{VSB}}(t)$ using synchronous demodulation at the receiver. This is done by multiplying the incoming VSB signal $\varphi_{\text{VSB}}(t)$ by $2\cos\omega_c t$. The product $e(t)$ is given by

$$e(t) = 2\varphi_{\text{VSB}}(t)\cos\omega_c t \Longleftrightarrow [\Phi_{\text{VSB}}(\omega+\omega_c) + \Phi_{\text{VSB}}(\omega-\omega_c)]$$

The signal $e(t)$ is further passed through the low-pass equalizer filter of transfer function $H_o(\omega)$. The output of the equalizer filter is required to be $m(t)$. Hence, the output signal spectrum is given by

$$M(\omega) = [\Phi_{\text{VSB}}(\omega+\omega_c) + \Phi_{\text{VSB}}(\omega-\omega_c)]H_o(\omega)$$

Substituting Eq. (4.18) into this equation and eliminating the spectra at $\pm 2\omega_c$ [suppressed by a low-pass filter $H_o(\omega)$], we obtain

$$M(\omega) = M(\omega)[H_i(\omega+\omega_c) + H_i(\omega-\omega_c)]H_o(\omega) \tag{4.19}$$

Hence[*]

$$H_o(\omega) = \frac{1}{H_i(\omega+\omega_c) + H_i(\omega-\omega_c)} \qquad |\omega| \le 2\pi B \tag{4.20}$$

Note that because $H_i(\omega)$ is a bandpass filter, the terms $H_i(\omega\pm\omega_c)$ contain low-pass components.

---

[*] If we choose $H_i(\omega)$ such that

$$H_i(\omega+\omega_c) + H_i(\omega-\omega_c) = 1 \qquad |\omega| \le 2\pi B \tag{4.21a}$$

The output filter is just a simple low-pass filter with transfer function $H_o(\omega) = 1$ over the baseband $|\omega| = 2\pi B$, because for a real filter, $H_i(-\omega) = H_i^*(\omega)$. Eq. (4.21a) can be expressed as

$$H_i(\omega_c+\omega) + H_i^*(\omega_c-\omega) = 1 \qquad |\omega| \le 2\pi B \tag{4.21b}$$

or

$$H_i(\omega_c+x) + H_i^*(\omega_c-x) = 1 \qquad |x| \le 2\pi B \tag{4.21c}$$

EXAMPLE 4.8 The carrier frequency of a certain VSB signal is $\omega_c = 20$ kHz, and the baseband signal bandwidth is 6 kHz. The VSB shaping filter $H_i(\omega)$ at the input, which cuts off the lower sideband gradually over 2 kHz, is shown in Fig. 4.23a. Find the output filter $H_o(\omega)$ required for distortionless reception.

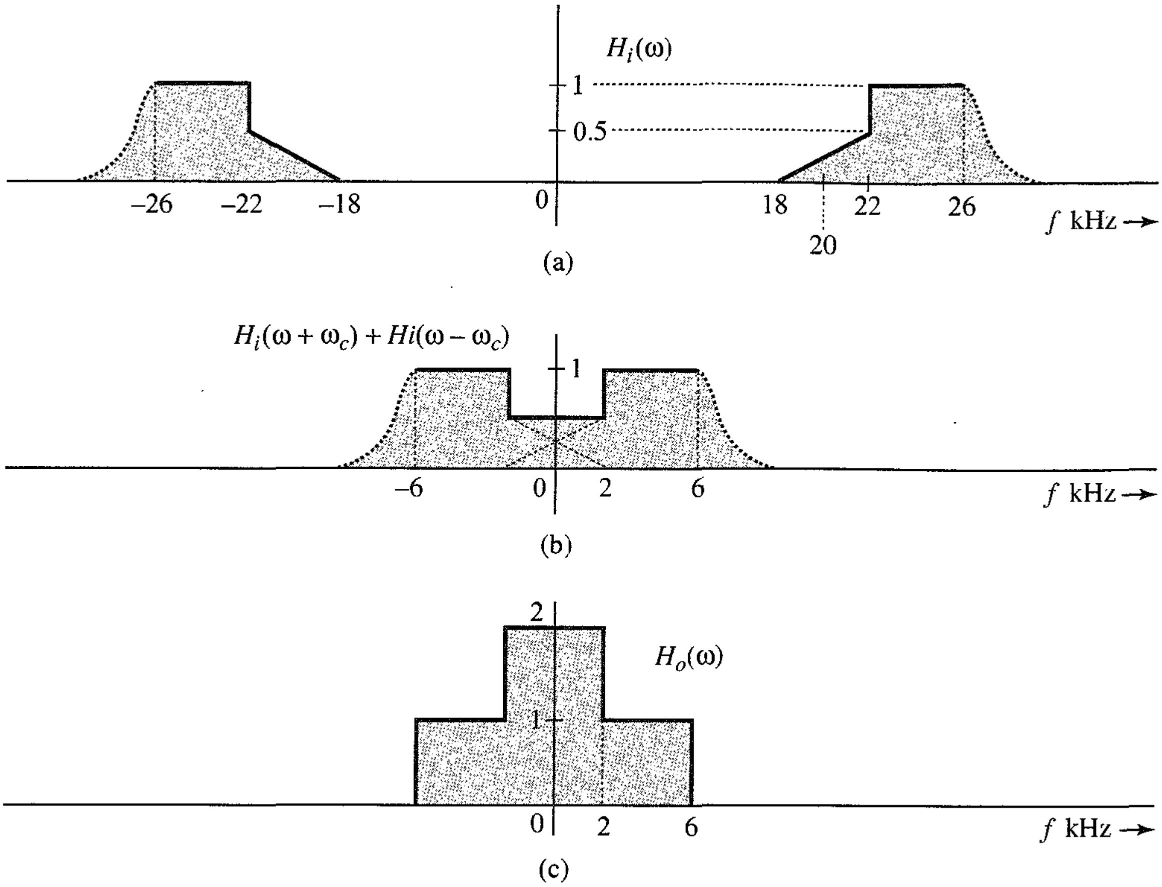


Figure 4.23 VSB out filter.

Figure 4.23b shows the low-pass segments of $H_i(\omega + \omega_c) + H_i(\omega - \omega_c)$. We are interested in this spectrum only over the baseband (the remaining undesired portion is suppressed by the output filter). This spectrum is 0.5 over the band of 0 to 2 kHz, and is 1 over 2 to 6 kHz, as shown in Fig. 4.23b. Figure 4.23c shows the desired output filter $H_o(\omega)$, which is the reciprocal of the spectrum in Fig. 4.23b [see Eq. (4.20)].

Envelope Detection of VSB+C Signals

That VSB+C signals can be envelope detected may be proved by using exactly the same argument used in proving the case for SSB+C signals. Both the SSB and the VSB modulated signals have the same form, with $m_h(t)$ in SSB replaced by some other signal $m_s(t)$ in VSB. This is because the VSB signal is a bandpass signal, which can be expressed in terms of the quadrature components in Eq. (3.38).

We have shown that SSB+C requires a much larger carrier than DSB+C (AM) for envelope detection. Because VSB+C is an in-between case, the added carrier required in VSB is larger than that in AM, but smaller than that in SSB+C.

Use of VSB in Broadcast Television

VSB is a clever compromise between SSB and DSB, which makes it very attractive for television broadcast systems. The baseband video signal of television occupies an enormous bandwidth of 4.5 MHz, and a DSB signal needs a bandwidth of 9 MHz. It would seem desirable to use SSB in order to conserve the bandwidth. Unfortunately, this creates several problems. First, the baseband video signal has sizable power in the low-frequency region, and consequently it is difficult to suppress one sideband completely. Second, for a broadcast receiver, an envelope detector is preferred over a synchronous one in order to reduce the receiver cost. We have seen earlier that SSB+C has a very low power efficiency. Moreover, use of SSB will increase the receiver cost.

The DSB spectrum of a television signal is shown in Fig. 4.24a. The vestigial shaping filter $H_i(\omega)$ cuts off the lower sideband spectrum gradually starting at 0.75 MHz to 1.25 MHz below the carrier frequency f_c, as shown in Fig. 4.24b. The receiver output filter $H_o(\omega)$ is designed according to Eq. (4.20). The resulting VSB spectrum bandwidth is 6 MHz. Compare this with the DSB bandwidth of 9 MHz and the SSB bandwidth of 4.5 MHz.

Linearity of Amplitude Modulation

In all the types of modulation discussed thus far, the modulated signal (excluding the carrier term) satisfies the principles of superposition. For example, if modulating signals $m_1(t)$ and $m_2(t)$ produce modulated signals* $\varphi_1(t)$ and $\varphi_2(t)$, respectively, then the modulating signal $k_1m_1(t) + k_2m_2(t)$ produces the modulated signal $k_1\varphi_1(t) + k_2\varphi_2(t)$. The reader can verify linearity for all types of amplitude modulation (DSB, SSB, AM, and VSB). This property is valuable in analysis. Because any signal can be expressed as a sum (discrete or in continuum)

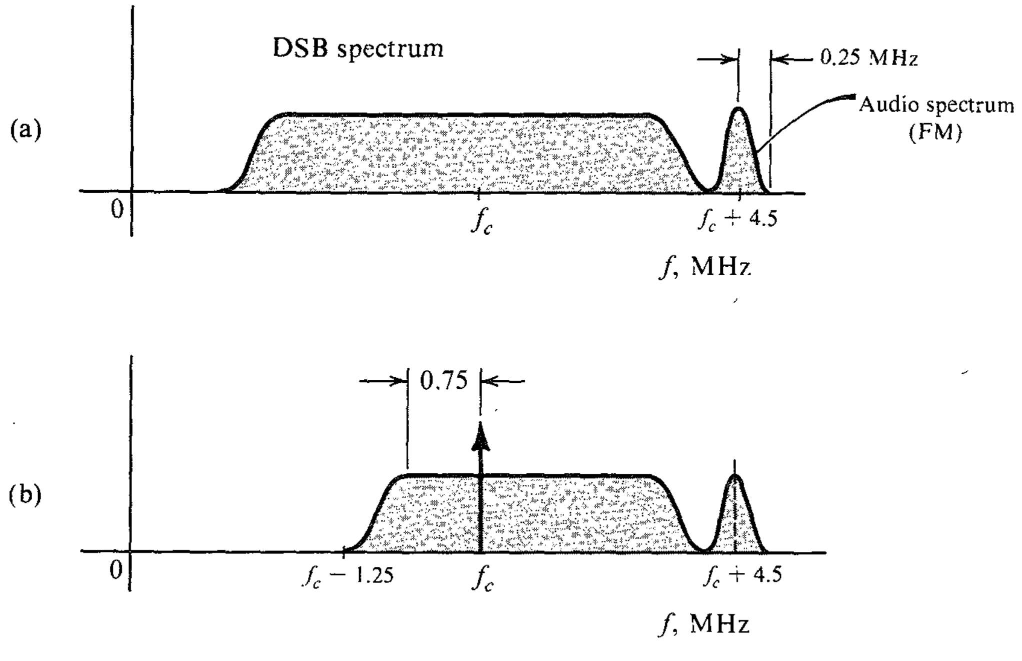


Figure 4.24 Television signal spectra. (a) DSB signal. (b) Signal transmitted.

---

* Note that we are excluding the carrier term from $\varphi_1(t)$ and $\varphi_2(t)$. In short, superposition applies to the suppressed carrier portion only. For more discussion, see Van Trees.[5]

of sinusoids, the complete description of the modulation system can be expressed in terms of tone modulation. For example, if $m(t) = \cos \omega_m t$ (tone modulation), the DSB-SC signal is

$$\cos \omega_m t \cos \omega_c t = \tfrac{1}{2}[\cos(\omega_c - \omega_m)t + \cos(\omega_c + \omega_m)t]$$

This shows that DSB-SC translates a frequency ω_m to two frequencies, $\omega_c - \omega_m$ (LSB) and $\omega_c + \omega_m$ (USB). We can generalize this result to any nonsinusoidal modulating signal $m(t)$. This is precisely the result obtained earlier by using a more general analysis.

4.7 CARRIER ACQUISITION

In the suppressed-carrier amplitude-modulated system (DSB-SC, SSB-SC, and VSB-SC), one must generate a local carrier at the receiver for the purpose of synchronous demodulation. Ideally, the local carrier must be in frequency and phase synchronism with the incoming carrier. Any discrepancy in the frequency or phase of the local carrier gives rise to distortion in the detector output.

Consider a DSB-SC case where a received signal is $m(t)\cos \omega_c t$ and the local carrier is $2\cos[(\omega_c + \Delta\omega)t + \delta]$. The local-carrier frequency and phase errors in this case are $\Delta\omega$ and δ, respectively. The product of the received signal and the local carrier is $e(t)$, given by

$$\begin{aligned} e(t) &= 2m(t)\cos \omega_c t \cos[(\omega_c + \Delta\omega)t + \delta] \\ &= m(t)\{\cos[(\Delta\omega)t + \delta] + \cos[(2\omega_c + \Delta\omega)t + \delta]\} \end{aligned} \tag{4.22}$$

The second term on the right-hand side is filtered out by the low-pass filter, leaving the output $e_o(t)$ as

$$e_o(t) = m(t)\cos[(\Delta\omega)t + \delta] \tag{4.23}$$

If $\Delta\omega$ and δ are both zero (no frequency or phase error), then

$$e_o(t) = m(t)$$

as expected. Let us consider two special cases. If $\Delta\omega = 0$, Eq. (4.23) reduces to

$$e_o(t) = m(t)\cos\delta \tag{4.24a}$$

This output is proportional to $m(t)$ when δ is a constant. The output is maximum when $\delta = 0$ and minimum (zero) when $\delta = \pm\pi/2$. Thus, the phase error in the local carrier causes the attenuation of the output signal without causing any distortion, as long as δ is constant. Unfortunately, the phase error δ may vary randomly with time. This may occur, for example, because of variations in the propagation path. This causes the gain factor $\cos\delta$ at the receiver to vary randomly and is undesirable.

Next we consider the case where $\delta = 0$ and $\Delta\omega \neq 0$. In this case, Eq. (4.23) becomes

$$e_o(t) = m(t)\cos(\Delta\omega)t \tag{4.24b}$$

The output here is not merely an attenuated replica of the original signal but is also distorted. Because $\Delta\omega$ is usually small, the output is the signal $m(t)$ multiplied by a low-frequency sinusoid. This causes the amplitude of the desired signal $m(t)$ to vary from maximum to zero periodically at twice the period of the beat frequency $\Delta\omega$. This "beating" effect is catastrophic even for a small frequency difference. The effect of this distortion even for a small frequency mismatch, say $\Delta f = 1$ Hz, is similar to the output when some restless kid is fiddling with its volume control knob up and down continuously twice a second.

To ensure identical carrier frequencies at the transmitter and the receiver, we can use quartz crystal oscillators, which generally are very stable. Identical crystals are cut to yield the same frequency at the transmitter and the receiver. At very high carrier frequencies, where the crystal dimensions become too small to match exactly, quartz-crystal performance may not be adequate. In such a case, a carrier, or **pilot**, is transmitted at a reduced level (usually about -20 dB) along with the sidebands. The pilot is separated at the receiver by a very narrow-band filter tuned to the pilot frequency. It is amplified and used to synchronize the local oscillator. The phase-locked loop (PLL), which plays an important role in carrier acquisition, will now be discussed.

The nature of the distortion caused by asynchronous carrier in SSB-SC is somewhat different than that in DSB-SC. In SSB-SC, when the carrier at the receiver is $2\cos[(\omega_c + \Delta\omega)t]$, the output is $m(t)$ with all its spectral components shifted (offset) by $\Delta\omega$ (see Prob. 4.5-5). Such a shift of every frequency component by a fixed amount $\Delta\omega$ destroys the harmonic relationship between frequency components. For instance, if $\Delta f = 10$ Hz, then the components of frequencies 1000 and 2000 Hz will be shifted to frequencies 1010 and 2010. This destroys their harmonic relationship. But unless Δf is very large, such a change does not destroy intelligibility of the output (as the beating effect does in the case of DSB-SC). For audio signals $\Delta f < 30$ Hz does not significantly affect the signal quality. $\Delta f > 30$ Hz results in a sound quality similar to that of Donald Duck. But the intelligibility is not completely lost.

When the carrier is $\cos(\omega_c t + \theta)$, the output is the signal $m(t)$ with the phases of all its spectral components shifted by θ (see Prob. 4.5-5). The phase distortion in SSB-SC also gives rise to the Donald Duck sound effect. This discussion shows that the problem of carrier synchronization is more critical in DSB-SC than in SSB-SC.

Phase-Locked Loop (PLL)

The **phase-locked loop (PLL)** can be used to track the phase and the frequency of the carrier component of an incoming signal. It is, therefore, a useful device for synchronous demodulation of AM signals with suppressed carrier or with a little carrier (the pilot). It can also be used for the demodulation of angle-modulated signals, especially under low SNR conditions. For this reason, the PLL is used in such applications as space-vehicle-to-earth data links, where there is a premium on transmitter weight, or where the loss along the transmission path is very large; and, more recently, in commercial FM receivers.

A PLL has three basic components:

1. A voltage-controlled oscillator (VCO)
2. A multiplier, serving as a phase detector (PD) or a phase comparator
3. A loop filter $H(s)$

The operation of the PLL is similar to that of a feedback system (Fig. 4.25a). In a typical feedback system, the signal fed back tends to follow the input signal. If the signal fed back is not equal to the input signal, the difference (known as the error) will change the signal fed back until it is close to the input signal. A PLL operates on a similar principle, except that the quantity fed back and compared is not the amplitude, but the phase. The VCO adjusts its own

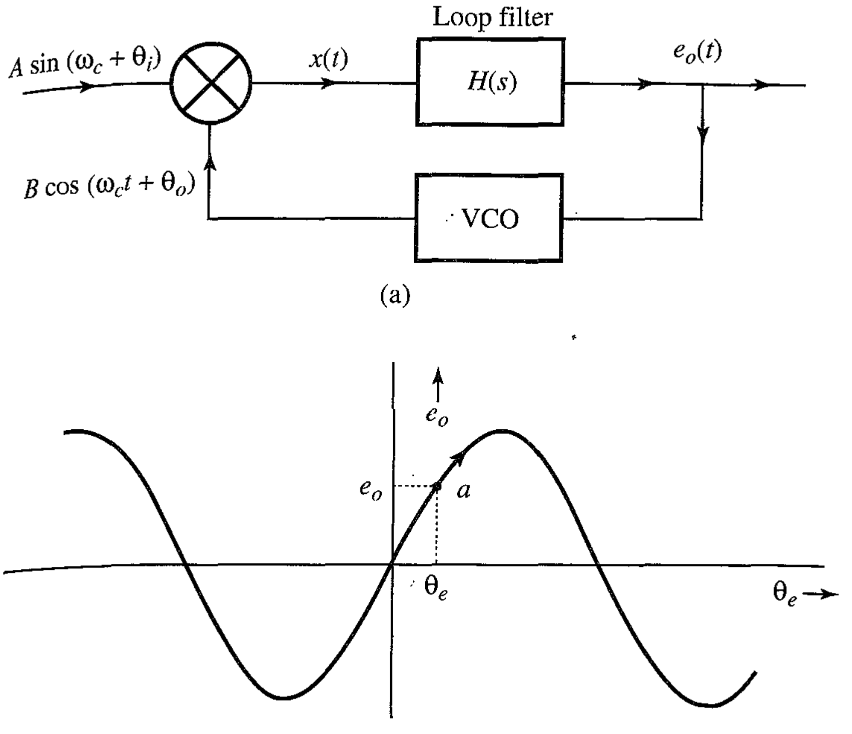


Figure 4.25 Phase-locked loop operation.

frequency until it is equal to that of the input sinusoid. At this point, the frequency and phase of the two signals are in synchronism (except for a possible difference of a constant phase).

Voltage-Controlled Oscillator (VCO): An oscillator whose frequency can be controlled by an external voltage is a **voltage-controlled oscillator (VCO)**. In a VCO, the oscillation frequency varies linearly with the input voltage. If a VCO input voltage is $e_o(t)$, its output is a sinusoid of frequency ω given by

$$\omega(t) = \omega_c + ce_o(t) \tag{4.25}$$

where c is a constant of the VCO and ω_c is the **free-running frequency** of the VCO [the VCO frequency when $e_o(t) = 0$]. The multiplier output is further low-pass-filtered by the loop filter and then applied to the input of the VCO. This voltage changes the frequency of the oscillator and keeps the loop **locked**.

How the PLL Works: Let the input to the PLL be $A\sin(\omega_c t + \theta_i)$, and let the VCO output be a sinusoid $B\cos(\omega_c t + \theta_o)$.[*] The multiplier output $x(t)$ is given by

$$x(t) = AB\sin(\omega_c t + \theta_i)\cos(\omega_c t + \theta_o) = \frac{AB}{2}[\sin(\theta_i - \theta_o) + \sin(2\omega_c t + \theta_i + \theta_o)]$$

The last term on the right-hand side, being a high-frequency signal, is suppressed by the loop filter, which is a low-pass narrow-band filter. Hence, $e_o(t)$, the input to the VCO, is given by

$$e_o = \frac{AB}{2}\sin\theta_e \qquad \theta_e = \theta_i - \theta_o \tag{4.26}$$

---

[*] It is not necessary for the VCO input and output frequencies to be equal. All that is needed is to set the VCO free-running frequency as close as possible to the incoming frequency. If the VCO output is $B\cos(\hat{\omega}_c t + \theta_o)$, we can express it as $B\cos(\omega_c t + \hat{\theta}_o)$, where $\hat{\theta}_o = [(\hat{\omega}_c - \omega_c)t + \theta_o]$.

where θ_e is the phase error $(\theta_i - \theta_o)$. Figure 4.25b shows the plot of e_o vs. θ_e. Using this plot, we can explain the tracking mechanism as follows.

Suppose that the loop is *locked*, meaning that the frequencies of both the input and the output sinusoids are identical. This means things are in the steady state, and θ_i, θ_o, and θ_e are constant. Figure 4.25b shows a typical operating point a and the corresponding values of e_o and θ_e on the e_o vs. θ_e plot. Suppose further that the input sinusoid frequency suddenly increases from ω_c to ω_c+k. This means the incoming signal is $A\cos[(\omega_c + k)t + \theta_i] = A\cos(\omega_c t + \hat{\theta}_i)$, where $\hat{\theta}_i = kt + \theta_i$. Thus, the increase in the incoming frequency causes θ_i to increase to $\theta_i + kt$, thereby increasing θ_e. The operating point a now shifts upward along the e_o vs. θ_e characteristic in Fig. 4.25b. This increases e_o, which, in turn, increases the frequency of the VCO output to match the increase in the input frequency. A similar reasoning shows that if the input sinusoid frequency decreases, the PLL output frequency will also decrease correspondingly. Thus, the PLL tracks the input sinusoid. The two signals are said to be mutually **phase coherent** or **in phase lock**. The VCO thus tracks the frequency and the phase of the incoming signal. A PLL can track the incoming frequency only over a finite range of frequency shift. This range is called the **hold-in** or **lock** range. Moreover, if initially the input and output frequencies are not close enough, the loop may not acquire lock. The frequency range over which the input will cause the loop to lock is called the **pull-in** or **capture** range. Also if the input frequency changes too rapidly, the loop may not lock.

Although we assumed θ_i and θ_o to be constants, the preceding analysis is also valid if these angles are varying slowly with time. It is clear that the angle θ_o tends to follow the input angle θ_i closely when the PLL tracks the input signal; the difference $\theta_e = \theta_i - \theta_o$ is either a constant or a small number $\rightarrow 0$.

If the input sinusoid is noisy, the PLL not only tracks the sinusoid, but also cleans it up. The PLL can also be used as an FM demodulator and frequency synthesizer. Frequency multipliers and dividers can also be built using PLL. The PLL, being a relatively inexpensive integrated circuit, has become one of the most frequently used communication circuits.

In space vehicles, because of the Doppler shift and the oscillator drift, the frequency of the received signal has a lot of uncertainty. The Doppler shift of the carrier itself could be as high as ±75 kHz, whereas the desired modulated signal band may be just 10 Hz. To receive such a signal by conventional receivers would require a filter of bandwidth 150 kHz, when the desired signal has a bandwidth of only 10 Hz. This would cause an undesirable increase in the noise received (by a factor of 15,000), because the noise power is proportional to the bandwidth. The PLL proves convenient here because it tracks the received frequency continuously, and the filter bandwidth required is only 10 Hz.

Being a nonlinear system, the detailed analysis of PLL is rather involved and beyond our scope. Complete analysis of two special cases is carred out in Chapter 5.

Carrier Acquisition in DSB-SC

We shall now discuss two methods of carrier regeneration at the receiver in DSB-SC: signal squaring and Costas loop.

Signal-Squaring Method: An outline of this scheme is given in Fig. 4.26. The incoming signal is squared and then passed through a narrow (high Q) bandpass filter tuned to $2\omega_c$. The output of this filter is the sinusoid $k\cos 2\omega_c t$, with some residual unwanted signal. This signal is applied to a PLL to obtain a cleaner sinusoid of twice the carrier frequency, which

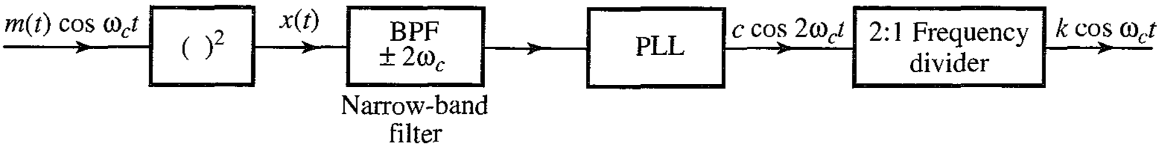


Figure 4.26 Generation of coherent demodulation carrier using signal squaring.

is passed through a 2:1 frequency divider to obtain a local carrier in phase and frequency synchronism with the incoming carrier. The analysis is straightforward. The squarer output $x(t)$ is

$$x(t) = [m(t)\cos\omega_c t]^2 = \tfrac{1}{2}m^2(t) + \tfrac{1}{2}m^2(t)\cos 2\omega_c t$$

Now $m^2(t)$ is a nonnegative signal, and therefore has a nonzero average value [in contrast to $m(t)$, which generally has a zero average value]. Let the average value, which is the dc component of $m^2(t)/2$, be k. We can now express $m^2(t)/2$ as

$$\frac{1}{2}m^2(t) = k + \phi(t)$$

where $\phi(t)$ is a zero mean baseband signal [$m^2(t)/2$ minus its dc component]. Thus,

$$\begin{aligned} x(t) &= \frac{1}{2}m^2(t) + \frac{1}{2}m^2(t)\cos 2\omega_c t \\ &= \frac{1}{2}m^2(t) + k\ \cos 2\omega_c t + \phi(t)\cos 2\omega_c t \end{aligned}$$

The bandpass filter is a narrow-band (high Q) filter tuned to frequency $2\omega_c$. It completely suppresses the signal $m^2(t)$, whose spectrum is centered at $\omega = 0$. It also suppresses most of the signal $\phi(t)\cos 2\omega_c t$. This is because although this signal spectrum is centered at $2\omega_c$, it has zero (infinitesimal) power at $2\omega_c$ since $\phi(t)$ has a zero dc value. Moreover this component is distributed over the band of $4B$ Hz centered at $2\omega_c$. Hence, very little of this signal passes through the narrow-band filter.* In contrast, the spectrum of $k\ \cos 2\omega_c t$ consists of impulses located at $\pm 2\omega_c$. Hence, all its power is concentrated at $2\omega_c$, and will pass through. Thus, the filter output is $k\ \cos 2\omega_c t$ plus a small undesired residue from $\phi(t)\cos 2\omega_c t$. This residue can be suppressed by using a PLL, which tracks $k\ \cos 2\omega_c t$. The PLL output, after passing through a 2:1 frequency divider, yields the desired carrier. One qualification is in order. Because the incoming signal sign is lost in the squarer, we have a sign ambiguity (or phase ambiguity of π) in the carrier generated. This is immaterial for analog signals. For a digital baseband signal, however, the carrier sign is essential, and this method, therefore, cannot be used directly.

Costas Loop: Yet another scheme, proposed by Costas,[6] for generating a local carrier, is shown in Fig. 4.27. The incoming signal is $m(t)\ \cos(\omega_c t + \theta_i)$. At the receiver, a VCO generates the carrier $\cos(\omega_c t + \theta_o)$. The phase error is $\theta_e = \theta_i - \theta_o$. Various signals are indicated in Fig. 4.27. The two low-pass filters suppress high-frequency terms to yield $m(t)\cos\theta_e$ and $m(t)\sin\theta_e$, respectively. These outputs are further multiplied to give $m^2(t)\ \sin 2\theta_e$. When this is passed through a narrow-band low-pass filter, the output is $k \sin 2\theta_e$, where k is the dc component of $m^2(t)/2$. The signal $k \sin 2\theta_e$ is applied to the input of a VCO with quiescent frequency ω_c. The input $k \sin 2\theta_e$ increases the output frequency, which, in turn, reduces θ_e. This mechanism was fully discussed earlier [see Eq. (4.26) and Fig. 4.25].

---

* This will also explain why we cannot extract the carrier directly from $m(t)\cos\omega_c t$ by passing it through a narrow-band filter centered at ω_c. The reason is that the power of $m(t)\cos\omega_c t$ at ω_c is zero because $m(t)$ has no dc component [the average value of $m(t)$ is zero].

Carrier Acquisition in SSB-SC

For the purpose of synchronization at the SSB receiver, one may use highly stable crystal oscillators, with crystals cut for the same frequency at the transmitter and the receiver. At very high frequencies, where even quartz crystals may have inadequate performance, a pilot carrier may be transmitted. These are the same methods used for DSB-SC. However, the received-signal squaring technique as well as the Costas loop used in DSB-SC cannot be used for SSB-SC. This can be seen by expressing the SSB signal as

$$\begin{aligned}\varphi_{\text{SSB}}(t) &= m(t)\cos\omega_c t \mp m_h(t)\sin\omega_c t \\ &= E(t)\cos\left[\omega_c t + \theta(t)\right]\end{aligned}$$

where

$$E(t) = \sqrt{m^2(t) + m_h^2(t)}$$

$$\theta(t) = \tan^{-1}\left[\frac{\pm m_h(t)}{m(t)}\right]$$

Squaring this signal yields

$$\begin{aligned}\varphi_{\text{SSB}}^2(t) &= E^2(t)\cos^2[\omega_c t + \theta(t)] \\ &= \frac{E^2(t)}{2}\{1 + \cos\left[2\omega_c t + 2\theta(t)\right]\}\end{aligned}$$

The signal $E^2(t)$ is eliminated by a bandpass filter. Unfortunately, the remaining signal is not a pure sinusoid of frequency $2\omega_c$ (as was the case for DSB). There is nothing we can do to remove

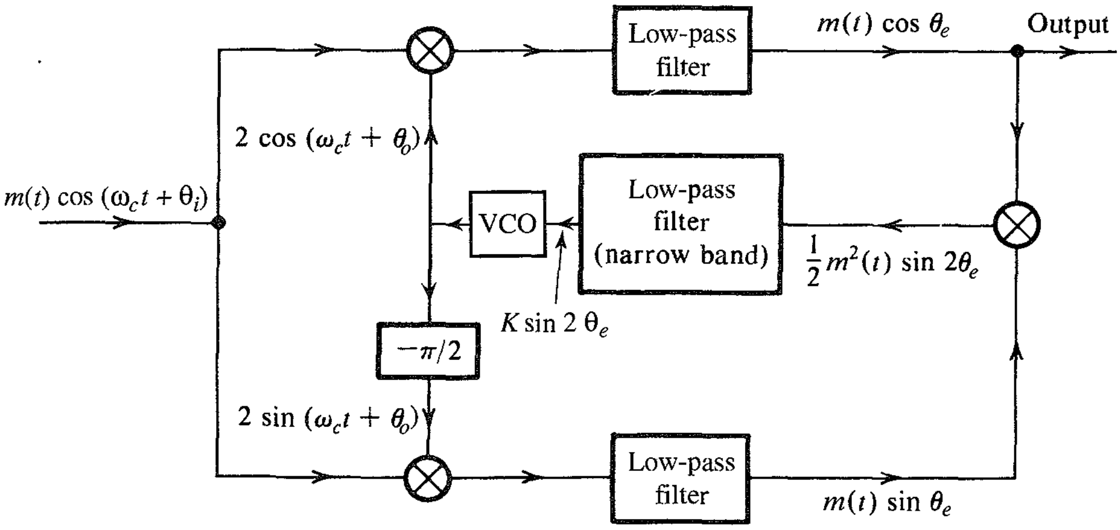

Low-pass filter
$m(t)$ cos θ_e
Output
2 cos $(\omega_c t + \theta_o)$
$m(t)$ cos $(\omega_c t + \theta_i)$
VCO
Low-pass filter (narrow band)
$\frac{1}{2}m^2(t)$ sin $2\theta_e$
K sin 2 θ_e
$-\pi/2$
2 sin $(\omega_c t + \theta_o)$
Low-pass filter
$m(t)$ sin θ_e

Figure 4.27 Costas phase-locked loop for the generation of a coherent demodulation carrier.

the time-varying phase $2\theta(t)$ from this sinusoid. Hence, for SSB, the squaring technique does not work. The same argument can be used to show that the Costas loop will not work either. These conclusions also apply to VSB signals.

Frequency-Division Multiplexing (FDM)

Signal multiplexing allows the transmission of several signals on the same channel. In Chapter 6, we shall discuss time-division multiplexing (TDM), where several signals time-share the same channel. In FDM, several signals share the band of a channel. Each signal is modulated by a different carrier frequency. The various carriers are adequately separated to avoid overlap (or interference) between the spectra of various modulated signals. These carriers are referred to as **subcarriers**. Each signal may use a different kind of modulation (for example, DSB-SC, AM, SSB-SC, VSB-SC, or even FM or PM). The modulated-signal spectra may be separated by a small guard band to avoid interference and facilitate signal separation at the receiver.

When all of the modulated spectra are added, we have a composite signal that may be considered as a baseband signal to further modulate a high-frequency [radio frequency (RF)] carrier for the purpose of transmission.

At the receiver, the incoming signal is first demodulated by the RF carrier to retrieve the composite baseband, which is then bandpass filtered to separate each modulated signal. Then each modulated signal is demodulated individually by an appropriate subcarrier to obtain all the basic baseband signals.

4.8 SUPERHETERODYNE AM RECEIVER

The radio receiver used in an AM system is called the **superheterodyne** AM receiver and is illustrated in Fig. 4.28. It consists of an RF (radio-frequency) section, a frequency converter (see Example 4.2), an intermediate-frequency (IF) amplifier, an envelope detector, and an audio amplifier.

The RF section is basically a tunable filter and an amplifier that picks up the desired station by tuning the filter to the right frequency band. The next section, the frequency mixer (converter), translates the carrier from ω_c to a fixed IF frequency of 455 kHz (see Example 4.2 for frequency conversion). For this purpose, it uses a local oscillator whose frequency f_{LO} is exactly 455 kHz above the incoming carrier frequency f_c; that is, $f_{\text{LO}} = f_c + f_{\text{IF}}$ ($f_{\text{IF}} =$ 455 kHz). Note that this is up-conversion. The tuning of the local oscillator and the RF tunable filter is done by one knob. Tuning capacitors in both circuits are ganged together and are designed so that the tuning frequency of the local oscillator is always 455 kHz above the tuning frequency of the RF filter. This means every station that is tuned in is translated to a fixed carrier frequency of 455 kHz by the frequency converter.

The reason for translating all stations to a fixed carrier frequency of 455 kHz is to obtain adequate selectivity. It is difficult to design sharp bandpass filters of bandwidth 10 kHz (the modulated audio spectrum) if the center frequency f_c is very high. This is particularly true if this filter is tunable. Hence, the RF filter cannot provide adequate selectivity against adjacent channels. But when this signal is translated to an IF frequency by a converter, it is further amplified by an IF amplifier (usually a three-stage amplifier), which does have good selectivity. This is because the IF frequency is reasonably low, and, second, its center frequency is fixed

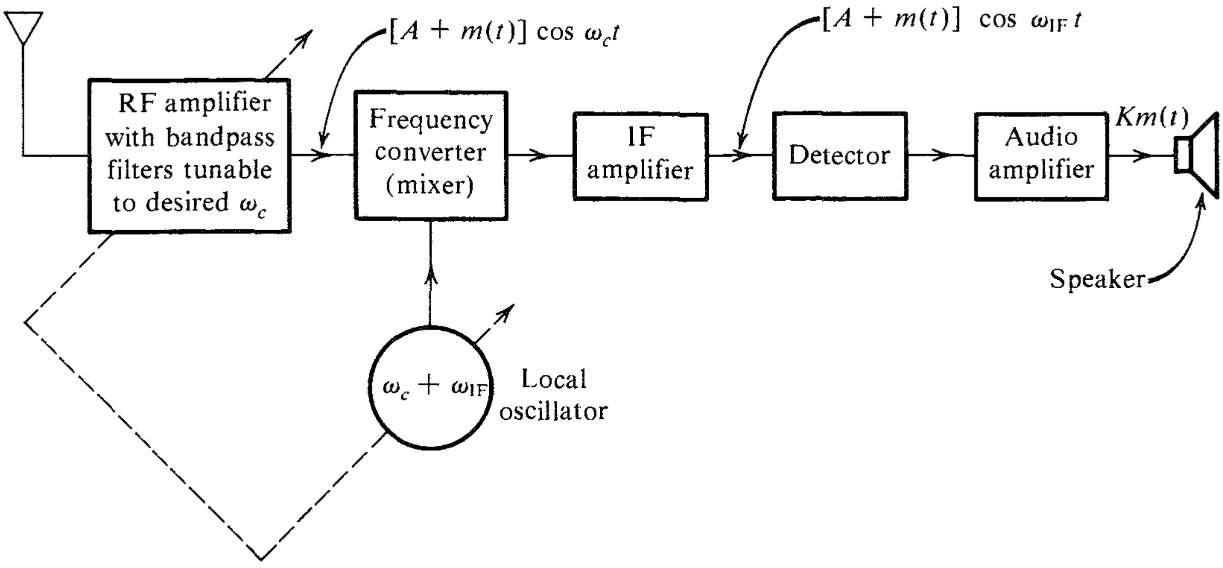

$[A + m(t)] \cos \omega_c t$
$[A + m(t)] \cos \omega_{IF} t$
RF amplifier with bandpass filters tunable to desired ω_c
Frequency converter (mixer)
IF amplifier
Detector
Audio amplifier
$Km(t)$
Speaker
$\omega_c + \omega_{IF}$
Local oscillator

Figure 4.28 Superheterodyne receiver.

and factory-tuned. Hence, the IF section can effectively suppress adjacent-channel interference because of its high selectivity. It also amplifies the signal for envelope detection.

In reality, practically all of the selectivity is realized in the IF section; the RF section plays a negligible role. The main function of the RF section is image frequency suppression. As observed in Example 4.2, the mixer, or converter, output consists of components of the difference between the incoming (f_c) and the local-oscillator (f_{LO}) frequencies (that is, $f_{IF} = |f_{LO} - f_c|$). Now, if the incoming carrier frequency $f_c = 1000$ kHz, then $f_{LO} = f_c + f_{RF} = 1000 + 455 = 1455$ kHz. But another carrier, with $f_c' = 1455 + 455 = 1910$ kHz, will also be picked up because the difference $f_c' - f_{LO}$ is also 455 kHz. The station at 1910 kHz is said to be the **image** of the station of 1000 kHz. Stations that are $2f_{IF} = 910$ kHz apart are called **image stations** and would both appear simultaneously at the IF output if it were not for the RF filter at receiver input. The RF filter may provide poor selectivity against adjacent stations separated by 10 kHz, but it can provide reasonable selectivity against a station separated by 910 kHz. Thus, when we wish to tune in a station at 1000 kHz, the RF filter, tuned to 1000 kHz, provides adequate suppression of the image station at 1910 kHz.

The receiver (Fig. 4.28) converts the incoming carrier frequency to the IF frequency by using a local oscillator of frequency f_{LO} higher than the incoming carrier frequency (up-conversion) and, hence, is called a superheterodyne receiver. The principle of superheterodyning, first introduce by E. H. Armstrong, is used in AM and FM as well as in television receivers. The reason for up-conversion rather than down-conversion is that the former leads to a smaller tuning range (smaller ratio of the maximum to minimum tuning frequency) for the local oscillator than does the latter. The broadcast-band frequencies range from 550 to 1600 kHz. The up-conversion f_{LO} ranges from 1005 to 2055 kHz (ratio of 2.045), whereas the down-conversion range of f_{LO} would be 95 to 1145 kHz (ratio of 12.05). It is much easier to design an oscillator that is tunable over a smaller frequency ratio.

The importance of the superheterodyne principle cannot be overstressed in radio and television broadcasting. In the early days (before 1919), the entire selectivity against adjacent stations was realized in the RF filter. Because this filter has poor selectivity, it was necessary to use several stages (several resonant circuits) in cascade for adequate selectivity. In the earlier receivers each filter was tuned individually. It was very time-consuming and cumbersome to tune in a station by bringing all resonant circuits into synchronism. This was improved upon as variable capacitors were ganged together by mounting them on the same shaft rotated by one knob. But variable capacitors are bulky, and there is a limit to the number that can be ganged together. This limited the selectivity available from receivers. Consequently, adjacent carrier frequencies had to be separated widely, resulting in fewer frequency bands. It was the superheterodyne receiver that made it possible to accommodate many more radio stations.

4.9 TELEVISION

In television, the central problem is the transmission of visual images by electrical signals. The image, or picture, can be thought of as a frame subdivided into several small squares, known as picture elements. A large number of picture elements in a given image means clearer reproduction (better resolution) at the receiver (see Fig. 4.29). The information of the entire picture is transmitted by transmitting an electrical signal proportional to the brightness level of the picture elements taken in a certain sequence. We start from the upper left-hand corner with element number 1 and scan the first row of elements (Fig. 4.30). Then we come back to the start of the second row, scan the second row, and continue this way until we finish the last row. The electrical signal thus generated during the entire scanning interval has the information of the picture.

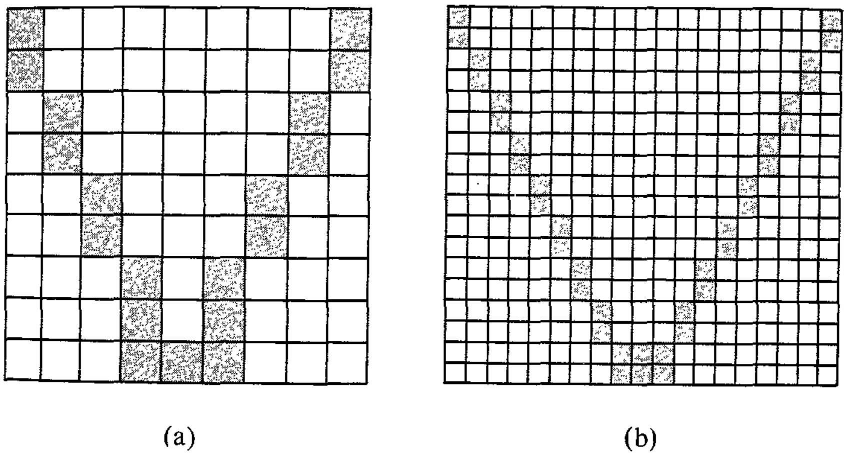


Figure 4.29 Effect of the number of picture elements on resolution.

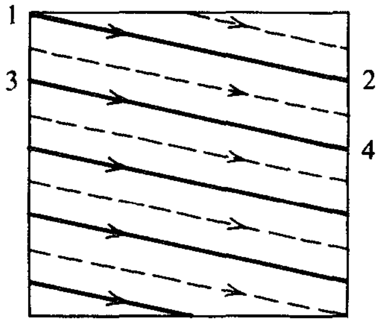


Figure 4.30 Scanning pattern (raster).

The image is furnished by the television camera tube. There exist a variety of camera tubes. The image orthicon is one example. In this tube, the optical system generates a focused image on a photo cathode, which eventually produces an electrically **charged image** on another surface, known as the **target mosaic**. What this means is that every point on the target-mosaic surface acquires a positive electric charge proportional to the brightness of the image. Thus, instead of a light image, we have a charge image. An electron gun now scans the target-mosaic surface with an electron beam in the manner shown in Fig. 4.30. The beam is controlled by a set of voltages across horizontal and vertical deflection plates. Periodic sawtooth signals (Fig. 4.31) are applied to these plates. The beam scans the horizontal line 1–2 in 53.5 μs and quickly flies back, in 10 μs, to the left to point 3 and scans line 3–4, and so on. On the target mosaic, where there is a high positive charge (corresponding to a higher brightness level), more electrons from the beam will be absorbed, and the return beam will have fewer electrons, giving a smaller current. Areas corresponding to darker elements (less positive charge) will return a large current. The scanning lines are not perfectly horizontal but have a small downward slope, because during the horizontal deflection the beam is also continuously deflected downward due to a slower vertical deflection signal (Fig. 4.31b). When all the horizontal lines are scanned, the vertical deflection signal goes to zero, which means the beam goes back to point 1 again and is ready to start the next frame.

Scanning is continuous at a rate of 60 picture frames per second. The electrical signal thus generated is a video signal corresponding to the visual image. This signal with some modifications (to be discussed later) VSB-modulates the video carrier of frequency f_c (see Fig. 4.24). This carrier is transmitted along with the frequency-modulated audio carrier of frequency f_α, which is 4.5 MHz higher than the video carrier frequency f_c, that is, $f_\alpha = f_c + 4.5$ MHz.

The receiver is similar to an oscilloscope. An electron gun with horizontal and vertical deflection plates generates an electron beam that scans the screen exactly in the same pattern and in synchronism with the scanning at the transmitter. When the electron beam flies back horizontally after completing each horizontal line, it will leave an unwanted flyback trace on the screen. To avoid this, a blanking pulse, known as the horizontal blanking pulse, is added during the flyback interval, which occurs at the end of each horizontal sweep. Similarly, a vertical blanking pulse is added at the end of each vertical sweep to eliminate the unwanted vertical retrace. These blanking pulses are added at the transmitter itself. We also need to add

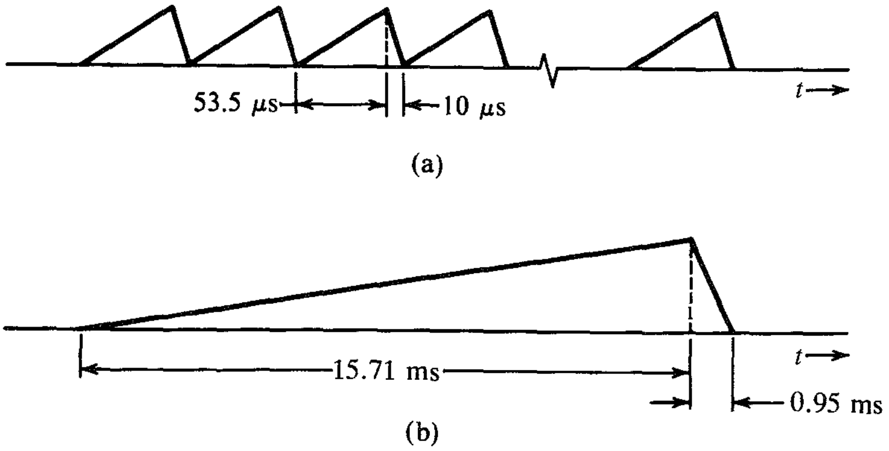


Figure 4.31 (a) Horizontal deflection signal. (b) Vertical deflection signal.

scan-synchronization information at the transmitter. This is done by adding a large pulse to each blanking pulse. A typical video signal is shown in Fig. 4.32. It is evident that over the entire flyback interval, the blanking pulse (at the black level) will eliminate the trace. Similarly, vertical blanking and synchronizing pulses, which are much wider than the corresponding horizontal pulses, are added to the video signal at the end of each vertical sweep. The video signal now VSB-modulates the carrier. We also add a carrier at this point (see Fig. 4.24). This VSB+C signal is transmitted along with the frequency-modulated audio signal. The transmitter block diagram is shown in Fig. 4.33a, the receiver block diagram in Fig. 4.33b. This is a superheterodyne receiver. The reasons for using a superheterodyne receiver were discussed earlier. The converter (a mixer) shifts the entire spectrum (video as well as frequency-modulated audio) to the IF frequency. This signal is now amplified and envelope detected. The audio signal is still of the frequency-modulated form with a carrier of 4.5 MHz. It is separated and demodulated. The video signal is amplified. Synchronizing pulses are separated and applied to the vertical and horizontal sweep generators. The video signal is clamped to the blanking pulses (dc restoration) and then applied to the picture tube.

Bandwidth Considerations

The number of horizontal lines used in the United States is 495 per frame. The time required for vertical retrace at the end of the scan is equivalent to that required for 30 horizontal lines. Hence, each frame is considered to have a total of 525 lines,* out of which only 495 are active. Images must be transmitted in a rapid succession of frames in order to create the illusion of continuity and avoid the flicker and jerky motion seen in old Charlie Chaplin movies. Because of the retinal property of retaining an image for a brief period even after the object is removed, it is necessary to transmit about 40 images, or frames, per second. In television we transmit only 30 frames per second in order to conserve bandwidth. To eliminate the flicker effect caused by the low frame rate, scanning the 495 lines is done in two successive patterns. In the first scanning pattern, called the first field, the entire image is scanned using only 247.5 lines (solid lines shown in Fig. 4.30). In the second scanning pattern, or second field, the image is

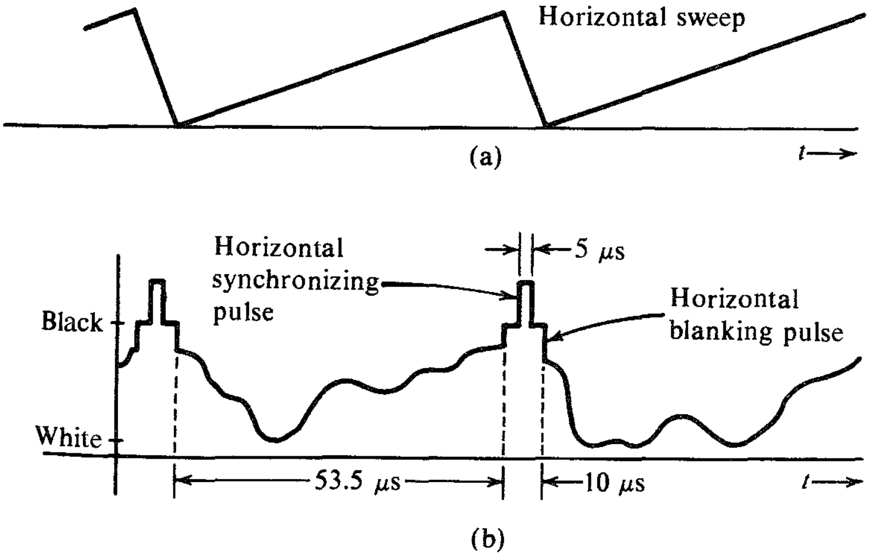
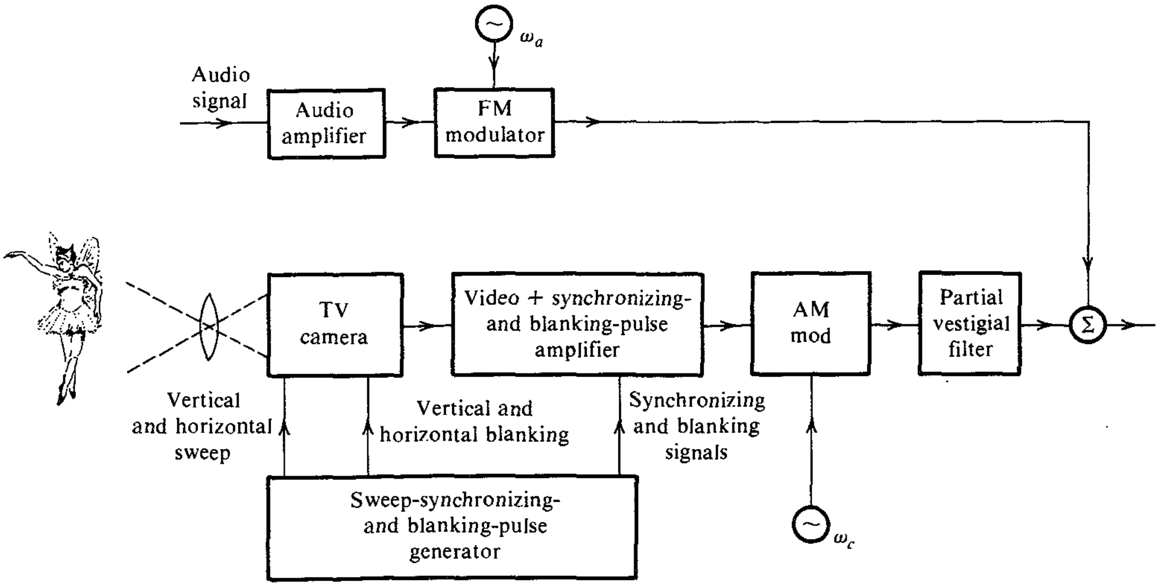


Figure 4.32 Television video signal.

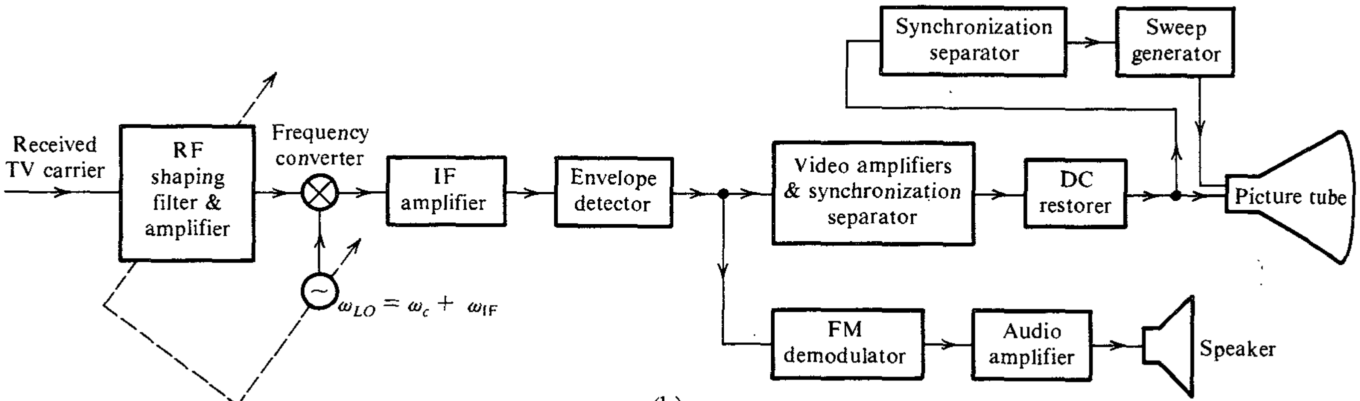
---

* In Europe, a total of 625 horizontal lines is used.


ω_a
Audio signal
Audio amplifier
FM modulator
TV camera
Video + synchronizing- and blanking-pulse amplifier
AM mod
Partial vestigial filter
Σ
Vertical and horizontal sweep
Vertical and horizontal blanking
Synchronizing and blanking signals
Sweep-synchronizing- and blanking-pulse generator
ω_c

(a)

Figure 4.33 (a) Television transmitter.


Received TV carrier
RF shaping filter & amplifier
Frequency converter
IF amplifier
Envelope detector
Synchronization separator
Sweep generator
Video amplifiers & synchronization separator
DC restorer
Picture tube
FM demodulator
Audio amplifier
Speaker
ω_LO = ω_c + ω_IF

(b)

Figure 4.33 (b) Television receiver.

scanned again by using 247.5 lines interlaced between lines of the first field (shown dotted in Fig. 4.30). The two fields together constitute a complete image, or frame. Thus, in reality there are only 30 complete frames per second, and a total equivalent of $525 \times 525 \times 30 = 8.27 \times 10^6$ picture elements per second.* We can estimate the transmission bandwidth of a video signal by observing that transmitting a video signal amounts to transmitting 8.27×10^6 pieces of information (or pulses) per second. Hence, the theoretical bandwidth required is half this, namely, 4.135 MHz (see Sec. 6.1.3).

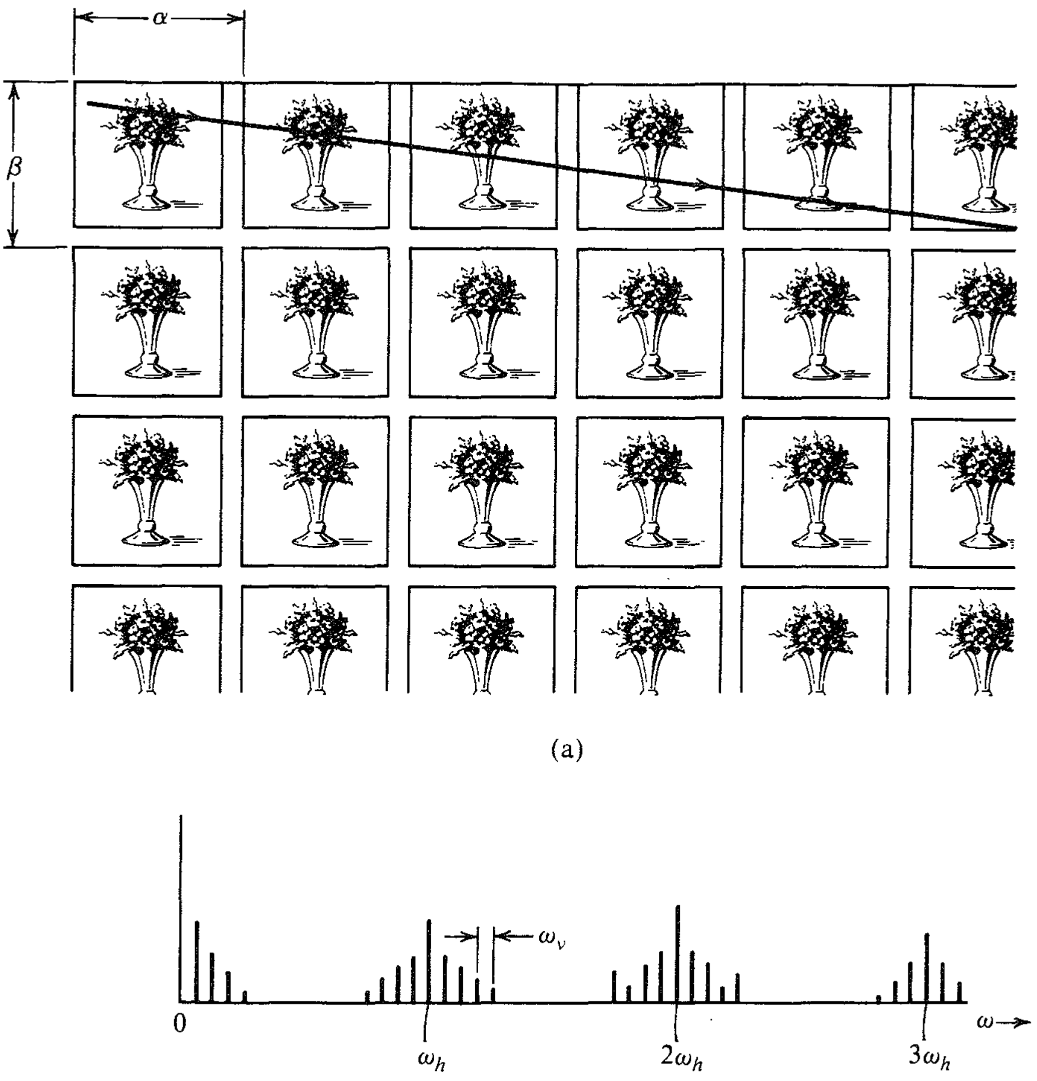
---

* Actually, the ratio of the image width to the image height (aspect ratio) is 4/3. Hence the number of picture elements will increase by a factor of 4/3. But this factor is almost canceled out because the scanning pattern does not align perfectly with the checkerboard pattern in Fig. 4.29, thus reducing the resolution by a factor of 0.70 (the Kell factor).

Video Spectrum

To begin with, consider a simple case of transmission of a still image. The scanning procedure discussed earlier is equivalent to scanning an array of the same image repeating itself in both dimensions, as shown in Fig. 4.34a. The brightness level b for this figure is a function of x (horizontal) and y (vertical) and can be expressed as $b(x, y)$. Because the picture repeats in the x as well as the y dimension, $b(x, y)$ is a periodic function of both x and y, with periods of α and β, respectively. Hence, $b(x, y)$ can be represented by a two-dimensional Fourier series with fundamental frequencies $2\pi/\alpha$ and $2\pi/\beta$, respectively,

$$b(x, y) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} B_{mn} \exp\left[j2\pi\left(\frac{mx}{\alpha} + \frac{ny}{\beta}\right)\right] \tag{4.27a}$$



(a)

(b)

Figure 4.34 (a) Model for scanning process using doubly periodic image fields. (b) Spectrum of the monochrome video signal.

If the scanning beam moves with a velocity v_x and v_y in the x and y directions, respectively, then $x = v_x t$ and $y = v_y t$, and the video signal $e(t)$ is

$$e(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} B_{mn} \exp\left[j2\pi\left(m\frac{v_x}{\alpha}t + n\frac{v_y}{\beta}t\right)\right] \tag{4.27b}$$

But α/v_x is the time required to scan one horizontal line, and β/v_y is the time required to scan the complete image,

$$\frac{\alpha}{v_x} = \frac{1}{30(525)} \quad \text{and} \quad \frac{\beta}{v_y} = \frac{1}{30}$$

and

$$e(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} B_{mn} \exp\left[j2\pi(15{,}750m + 30n)t\right] \tag{4.27c}$$

The video signal is periodic with fundamentals $f_h = 15.75$ kHz (horizontal-sweep frequency) and $f_v = 30$ Hz. The harmonics are spaced at 15.75-kHz intervals, and around each harmonic is clustered a satellite of harmonics 30 Hz apart, as shown in Fig. 4.34b.

This spectrum was derived for still-picture transmission. When motion or change occurs from frame to frame, $b(x, y)$ will not be periodic, and the spectrum will not be a line spectrum, but will have spreading or smearing. But empty spaces still exist between harmonics of f_h (15.75 kHz). We take advantage of these gaps to transmit the additional information of a color television signal over the same bandwidth.

The FCC allows a 6-MHz bandwidth for television broadcasting, with the frequency allocations as shown in Table 4.1.

Table 4.1
TV Channel Frequency Assignments

Channel Number	Frequency Band, MHz
VHF 2, 3, 4	54–72
VHF 5, 6	76–88
VHF 7–13	174–216
UHF 14–83	470–890

Compatible Color Television (CCTV)

All colors can be synthesized by mixing the three primary colors—blue, yellow, and red—in the right amounts. In television, blue, green (the combination of blue and yellow), and red are used instead for the practical reason of the availability of phosphors that glow with these colors when excited by an electron beam.

In color television cameras, the optical system resolves the image into three primary color (red, green, and blue) images. A set of three camera tubes produces three video signals $m_r(t)$, $m_g(t)$, and $m_b(t)$ from these images. We could transmit the three video signals and synthesize the color image at the receiver from the three signals. This, however, causes two difficulties. It requires three times as much bandwidth as that of monochrome (black-and-white) television, and, second, it is not compatible with the existing monochrome system because a monochrome television will receive only one of the primary colors.

These problems are solved by using signal matrixing. The information about $m_r(t)$, $m_g(t)$, and $m_b(t)$ can be transmitted by three signals, each being a linear combination of $m_r(t)$, $m_g(t)$, and $m_b(t)$, provided the three combinations are linearly independent. Thus, we can transmit the signals $m_L(t)$, $m_I(t)$, and $m_Q(t)$ given by

$$m_L(t) = 0.30m_r(t) + 0.59m_g(t) + 0.11m_b(t)$$

$$m_I(t) = 0.60m_r(t) + 0.28m_g(t) - 0.32m_b(t)$$

$$m_Q(t) = 0.21m_r(t) - 0.52m_g(t) + 0.31m_b(t)$$

Signals $m_r(t)$, $m_g(t)$, and $m_b(t)$ are normalized to a maximum value of 1 so that the amplitudes of each of these signals lie in the range of 0 to 1. Hence, $m_L(t)$ is always positive, whereas $m_I(t)$ and $m_Q(t)$ are bipolar. The signal $m_L(t)$ is known as the **luminance** signal because it has been found that this particular combination of the three primary-color signals closely matches the luminance of the conventional monochrome video signal. Hence, a black-and-white set need use only this signal for its operation.

The signals $m_I(t)$ and $m_Q(t)$ are known as the **chrominance** signals.* We could have chosen some other combinations instead of $m_I(t)$ and $m_Q(t)$. But these particular combinations are chosen because they use certain features of human color vision efficiently,[7] as explained next.

Multiplexing Luminance and Chrominance Signals: The luminance signal $m_L(t)$ is transmitted as a monochrome video signal occupying a bandwidth of 4.2 MHz. The chrominance signals $m_I(t)$ and $m_Q(t)$ also have the same bandwidth (namely, 4.2 MHz each). Subjective tests have shown, however, that the human eye is not perceptive to changes in chrominance (hue and saturation) over smaller areas. This means we can cut out high-frequency components without affecting the quality of the picture, because the eye would not have perceived them anyway. This enables us to limit the bandwidths of the $m_I(t)$ and $m_Q(t)$ to 1.6 and 0.6 MHz, respectively. The signal $m_I(t)$ is further split into two components, $m_{IH}(t)$ and $m_I(t) - m_{IH}(t)$. The high frequency component $m_{IH}(t)$ consists of the components of $m_I(t)$ in the range of 0.6 to 1.6 MHz. The remaining low-frequency component $m_I(t) - m_{IH}(t)$ consists of all the spectral components of $m_I(t)$ in the range of 0 to 0.6 MHz. Signals $m_Q(t)$ and $m_I(t) - m_{IH}(t)$ are sent by QAM, whereas $m_{IH}(t)$ is sent by LSB (Figs. 4.35 and 4.36). The subcarrier has frequency† $f_{cc} = 3.583125$ MHz. These spectra are generated as shown in Figs. 4.35. We generate the DSB-SC spectrum of $m_I(t)$. This spectrum is the sum of the DSB-SC spectra of both its components $m_{IH}(t)$ and $m_I(t) - m_{IH}(t)$. This spectrum occupies a band of 2 to 5.2 MHz. However, the bandpass filter (2 to 4.2 MHz) suppresses the USB portion of $m_{IH}(t)$, leaving only the LSB spectrum of $m_{IH}(t)$ (Fig. 4.36). We still have the DSB-SC spectrum for $m_I(t) - m_{IH}(t)$. Thus, $x_I(t)$ consists of an LSB for $m_{IH}(t)$ and a DSB-SC for $m_I(t) - m_{IH}(t)$, and can be expressed as

---

* These signals have an interesting interpretation in terms of the **hue** and the **saturation** of colors. Hue refers to the color, such as red, yellow, green, blue, or any color in between. Saturation, or color intensity, refers to the purity of the color. For example, a deep red has 100% saturation, but pink—which is a dilution of red with white—will have a lesser amount of saturation. Saturation is given by $\sqrt{m_I^2(t) + m_Q^2(t)}$, and hue is given by an angle $\tan^{-1}[m_Q(t)/m_I(t)]$. Each color has a certain hue, or angle. For example, red, blue, and green are at angles of 19°, 136°, and 242°, respectively.

† $f_{cc} = 227.5f_h = 227.5 \times 15.75$ kHz $= 3.583125$ MHz. Thus, f_{cc} lies midway between $227f_h$ and $228f_h$. This causes the chrominance signals' spectra to be shifted to gaps midway between harmonics of f_h (Fig. 4.36d). In practice, f_{cc} is made slightly smaller than $227.5f_h$ ($f_{cc} = 3.579545$ MHz) to avoid an objectionable beat frequency with the audio carrier,[7] which lies 4.5 MHz above the picture carrier. Because $f_h = f_{cc}/227.5$, $f_h = 15.7326$ kHz, and the field repetition frequency is actually 59.94 rather than 60.

$$\begin{aligned} x_I(t) &= \underbrace{[m_I(t) - m_{IH}(t)]\cos\omega_{cc}t}_{\text{DSB(QAM) for } m_I(t)-m_{IH}(t)} + \underbrace{m_{IH}(t)\cos\omega_{cc}t + m_{IH_h}(t)\sin\omega_{cc}t}_{\text{LSB for } m_{IH}(t)} \\ &= m_I(t)\cos\omega_{cc}t + m_{IH_h}(t)\sin\omega_{cc}t \end{aligned}$$

Moreover, the signal $x_Q(t) = m_Q(t)\sin\omega_{cc}(t)$. Hence, the composite multiplexed signal $m_c(t)$ is

$$m_c(t) = m_L(t) + m_Q(t)\sin\omega_{cc}t + m_I(t)\cos\omega_{cc}t + m_{IH_h}(t)\sin\omega_{cc}t$$

In addition, a sample of the subcarrier (color burst) is added to this multiplexed signal for frequency and phase synchronization of the locally generated subcarrier at the receiver. The color burst is added on the trailing edge of the horizontal blanking pulse. This composite video signal is now sent by VSB+C, as discussed in Sec. 4.6.

The Receiver: Because the CCTV system is required to be compatible with monochrome receivers, let us see what happens if we apply the signal $m_v(t)$ to a monochrome

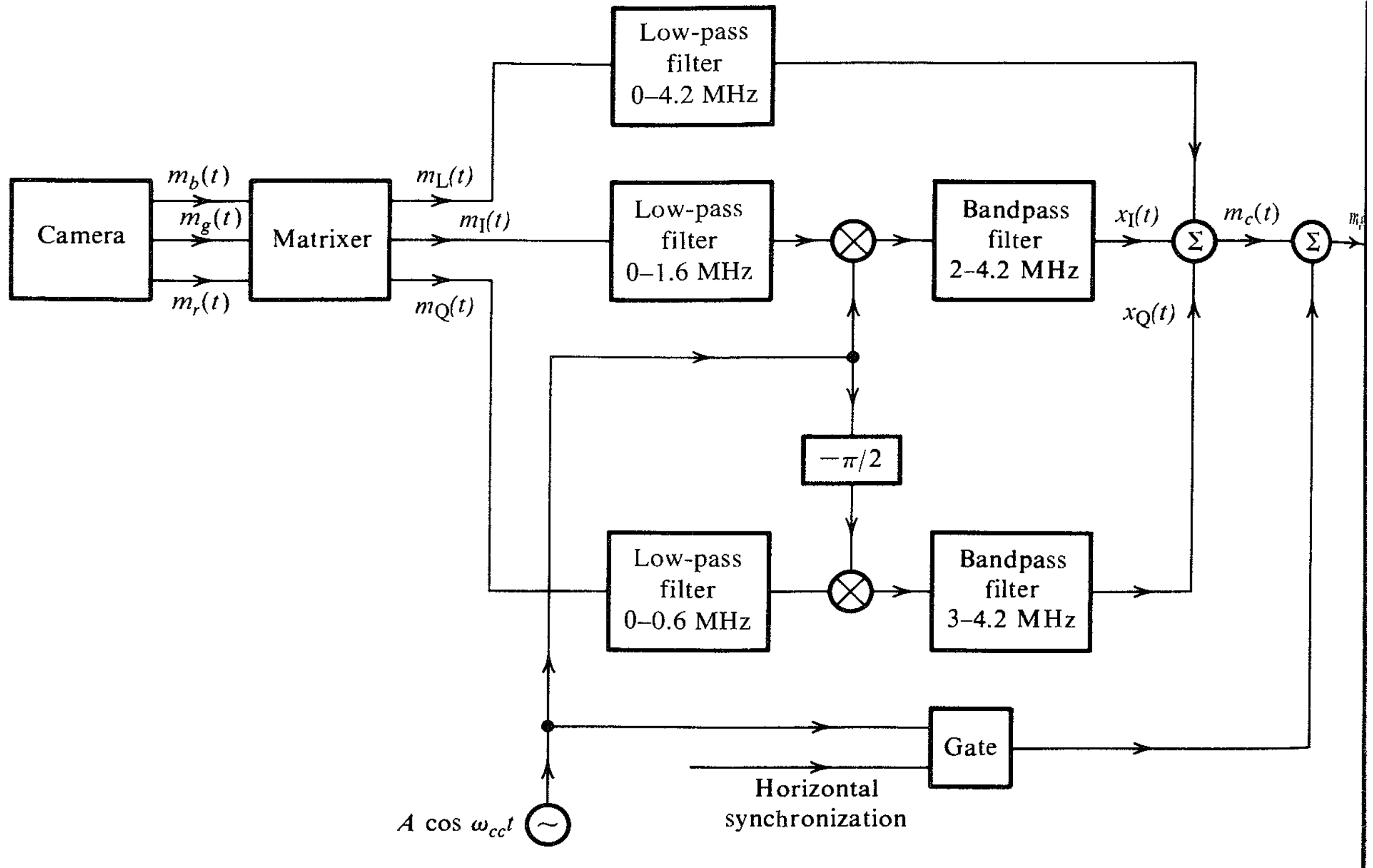

Low-pass filter 0–4.2 MHz
Camera
$m_b(t)$
$m_g(t)$
$m_r(t)$
Matrixer
$m_L(t)$
$m_I(t)$
$m_Q(t)$
Low-pass filter 0–1.6 MHz
Bandpass filter 2–4.2 MHz
$x_I(t)$
$m_c(t)$
Σ
Σ
$x_Q(t)$
$-\pi/2$
Low-pass filter 0–0.6 MHz
Bandpass filter 3–4.2 MHz
Gate
Horizontal synchronization
$A\cos\omega_{cc}t$

Figure 4.35 Multiplexing luminance and chrominance signals.

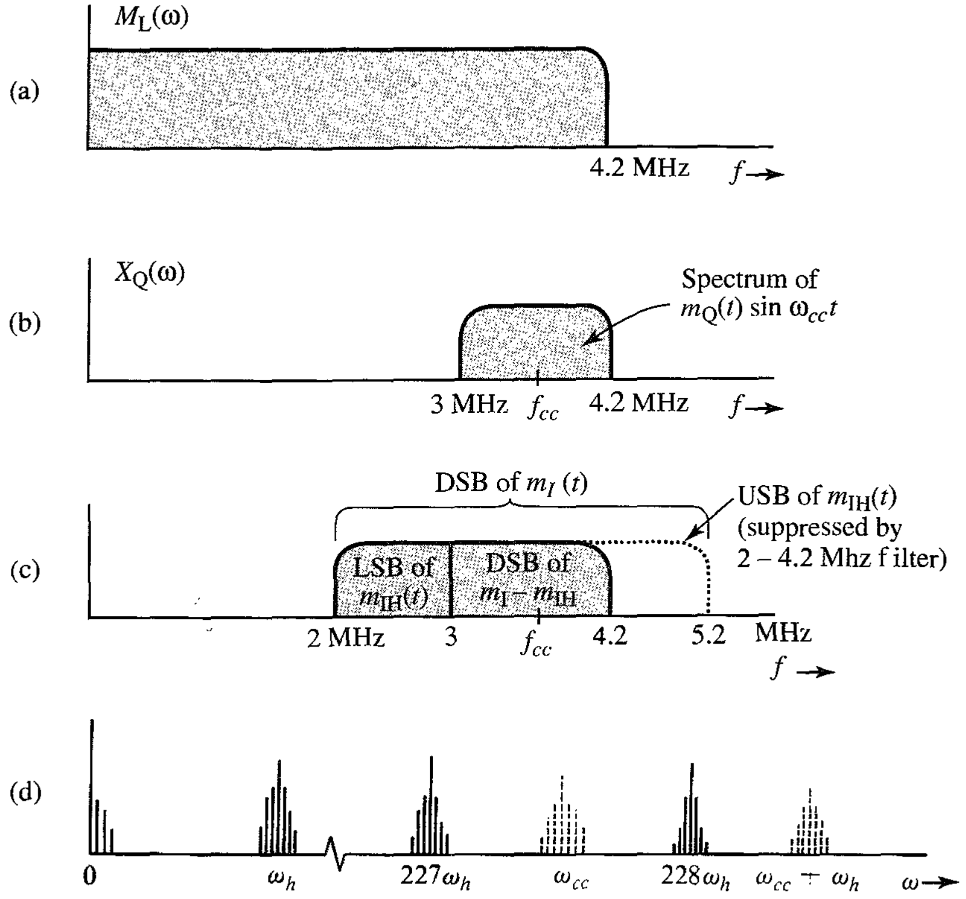


Figure 4.36 (a) Band occupied by $m_L(t)$. (b) Band occupied by $m_Q(t) \sin \omega_{cc}t$. (c) Bands occupied by LSB of $m_{IH}(t)$ and DSB of $m_I(t) - m_{IH}(t)$. (d) Interleaving of the chrominance and luminance signal spectra.

receiver. It may seem necessary to remove the chrominance signals from $m_v(t)$ before applying the signal to the picture tube. Fortunately, this is not necessary, because the interference of the chrominance signals with the luminance signal, although present on the screen, is practically invisible to the human eye. This happens because of the way chrominance signals are interlaced in the frequency domain and because of the persistence of human vision that tends to average out brightness over time as well as space.

The chrominance signal is superimposed on the luminance signal. Figure 4.37 shows the nature of the chrominance signals. Recall that $\omega_{cc} = 227.5\omega_h$. During one horizontal line, there will be 227.5 chrominance signal cycles. Hence, the chrominance signal changes continuously from positive to negative and vice versa in the horizontal direction. In addition, because there are 227.5 cycles in one line, if a chrominance signal begins with a positive cycle in the beginning of a line, it will end with a positive cycle at the end of the line. The next horizontal line will begin with a negative cycle of the chrominance signal (Fig. 4.37). Hence, in any given frame, the chrominance signal not only reverses its phase along the horizontal (x) direction but also reverses its phase along the vertical (y) direction (on the next horizontal line). But this is not all. During one field, the chrominance signal completes 227.5 x 525 cycles, and it returns to a given spot during the next field with opposite polarity. Hence, the chrominance signals reverse phase spatially (in the vertical and horizontal directions) as well as temporally at any given spot. Because the human eye is not sensitive to rapid time variations or rapid space variations, it can notice only space and time averages. This makes the chrominance signals practically invisible to the human eye. Thus the color signal is compatible with an unmodified monochrome receiver.

Demultiplexing: In a color receiver the received signal is demodulated exactly as in the monochrome case. This yields $m_v(t)$. This signal must now be demultiplexed to separate $m_L(t)$, $m_I(t)$, and $m_Q(t)$. The demultiplexing is shown in Fig. 4.38. The output of the 4.2-MHz filter contains $m_L(t)$, as well as modulated $m_I(t)$ and $m_Q(t)$ (Fig. 4.36). Because of the frequency interlacing discussed earlier, however, these signals are practically invisible. Hence, the output of the 4.2-MHz filter serves the function of $m_L(t)$. Next we demodulate $m_v(t)$ using carriers in phase quadrature. To determine the various signals in Fig. 4.38, we observe that the signal $z(t)$ in Fig. 4.38 consists of modulated $m_I(t)$ and $m_Q(t)$, plus the part of $m_L(t)$ in the band of 2 to 4.2 MHz. Let us denote this high-frequency component of $m_L(t)$ by $m_{LH}(t)$. Then,

$$z(t) = m_{LH}(t) + m_Q(t)\sin\omega_{cc}t + m_I(t)\cos\omega_{cc}t + m_{IH_h}(t)\sin\omega_{cc}t$$

Hence,

$$x_1(t) = 2m_{LH}(t)\cos\omega_{cc}t + m_Q(t)\sin 2\omega_{cc}t + m_I(t)(1+\cos 2\omega_c t) + m_{IH_h}(t)\sin 2\omega_{cc}t$$

The double-frequency terms will be suppressed by the bandpass filter. In addition, the signal $2m_{LH}(t)\cos\omega_{cc}t$ will be invisible because of the frequency-interlacing effect. This is because the spectrum of this signal is the spectrum of $m_L(t)$ shifted to $\omega_{cc} = 227.5\omega_h$, and it will become invisible because of the frequency interlacing discussed earlier. Hence, the filter output of the 0- to 1.6-MHz filter yields $m_I(t)$. Similarly, the output of the 0- to 0.6-MHz filter* yields

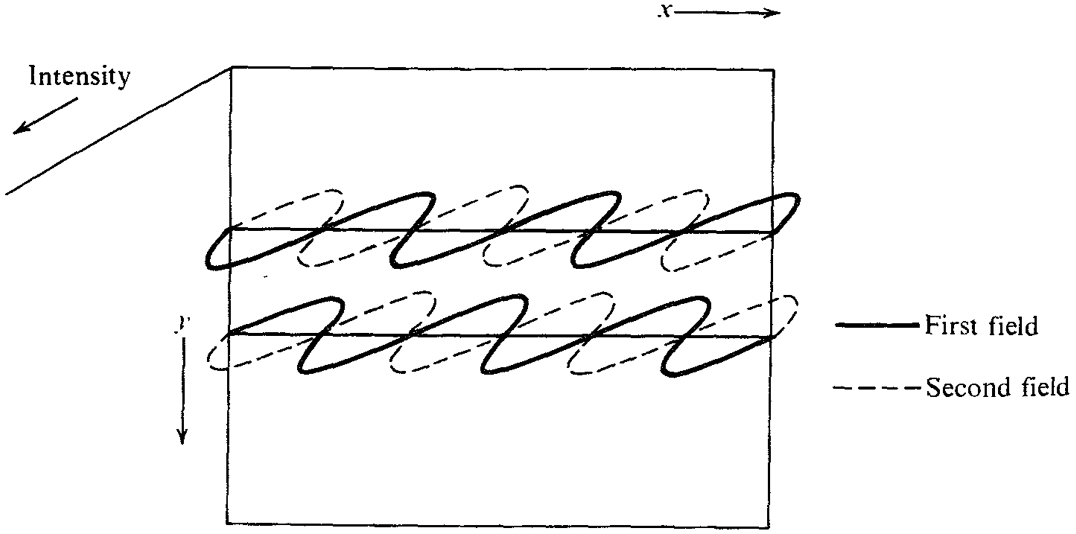


Figure 4.37 Temporal and spatial phase reversals of chrominance signals.

---

* This filter will suppress $m_{IH}(t)$, whose components lie in the range of 0.6 to 1.6 MHz.

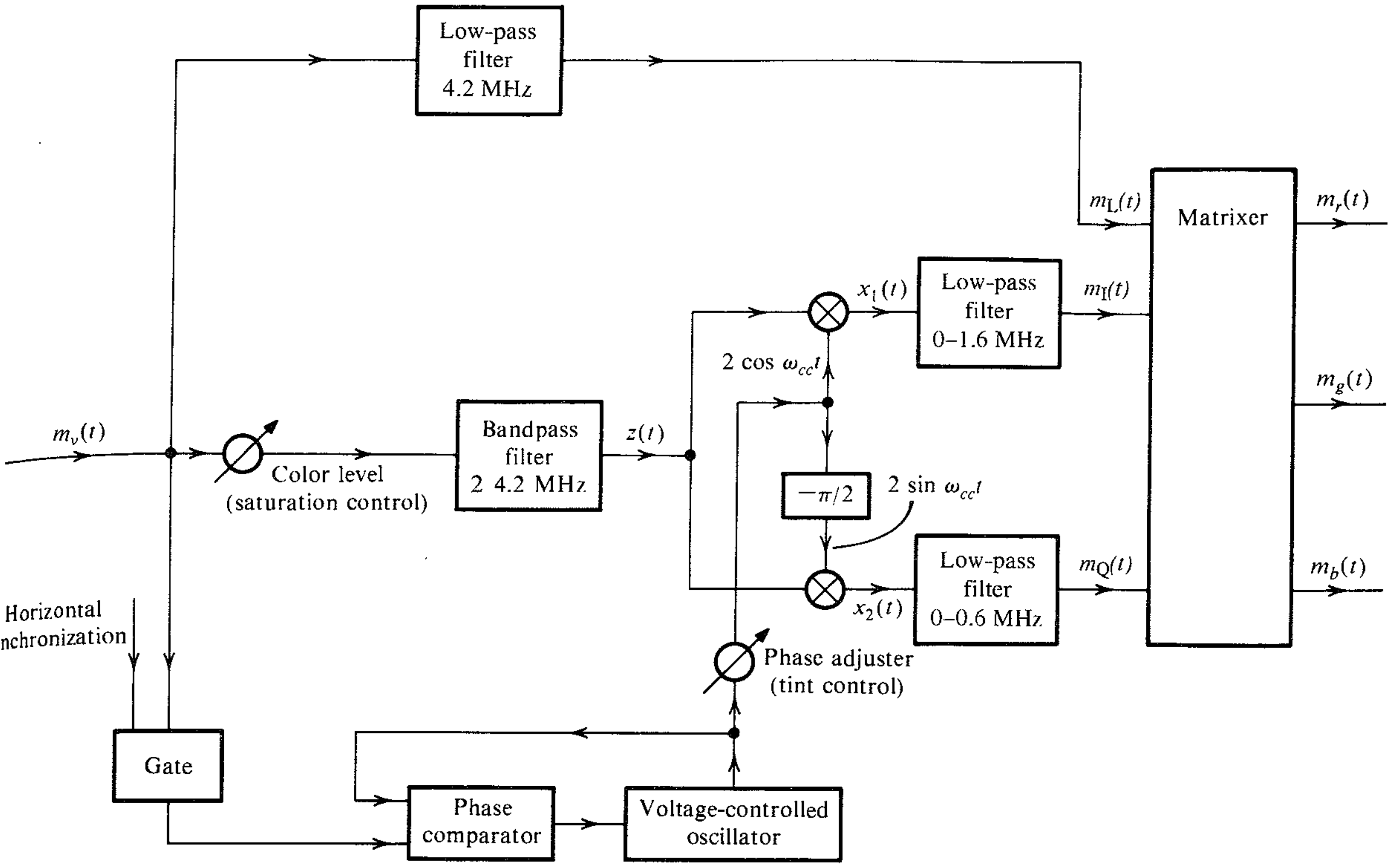

Low-pass filter 4.2 MHz
$m_L(t)$
Matrixer
$m_r(t)$
$x_1(t)$
Low-pass filter 0–1.6 MHz
$m_I(t)$
2 cos $\omega_{cc}t$
$m_g(t)$
$m_v(t)$
Color level (saturation control)
Bandpass filter 2 4.2 MHz
$z(t)$
$-\pi/2$
2 sin $\omega_{cc}t$
Low-pass filter 0–0.6 MHz
$m_Q(t)$
$m_b(t)$
$x_2(t)$
Horizontal synchronization
Phase adjuster (tint control)
Gate
Phase comparator
Voltage-controlled oscillator

Figure 4.38 Color television receiver.

$m_Q(t)$. The three signals $m_L(t)$, $m_I(t)$, and $m_Q(t)$ are then matrixed to obtain $m_r(t)$, $m_g(t)$, and $m_b(t)$.

A color carrier is generated using PLL. For this purpose we separate the color burst (Fig. 4.38) and apply it to a PLL whose output is the locally generated carrier that tracks the color burst. The phase of the locally generated carrier is adjustable. This is called **tint control**.

REFERENCES

1. D. H. Sheingold, ed., *Nonlinear Circuits Handbook*, Analog Devices, Inc., Norwood, MA, 1974.
2. Single Sideband Issue, *Proc. IRE*, vol. 44, Dec. 1956.
3. D. K. Weaver, Jr., "A Third Method of Generation and Detection of Single Sideband Signals," *Proc. IRE*, vol. 44, pp. 1703–1705, Dec. 1956.
4. Bell Telephone Laboratories, *Transmission Systems for Communication*, 4th ed., Murray Hill, NJ, 1970.
5. H. L. Van Trees, *Detection, Estimation, and Modulation Theory* (Part 1), Wiley, New York, 1968, Chapter 6.
6. J. P. Costas, "Synchronous Communication," *Proc. IRE*, vol. 44, pp. 1713–1718, Dec. 1956.

6. J. P. Costas, "Synchronous Communication," *Proc. IRE*, vol. 44, pp. 1713–1718, Dec. 1956.
7. L. H. Hansen, *Introduction to Solid-State Television Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

PROBLEMS

4.2-1 For each of the following baseband signals: (i) $m(t) = \cos 1000t$; (ii) $m(t) = 2\cos 1000t + \cos 2000t$; (iii) $m(t) = \cos 1000t \cos 3000t$:

(a) Sketch the spectrum of $m(t)$.

(b) Sketch the spectrum of the DSB-SC signal $m(t) \cos 10{,}000t$.

(c) Identify the upper sideband (USB) and the lower sideband (LSB) spectra.

(d) Identify the frequencies in the baseband, and the corresponding frequencies in the DSB-SC, USB, and LSB spectra. Explain the nature of frequency shifting in each case.

4.2-2 Repeat Prob. 4.2-1 [parts (**a**), (**b**), and (**c**) only] if: (**i**) $m(t) = \text{sinc}\,(100t)$; (**ii**) $m(t) = e^{-|t|}$; (**iii**) $m(t) = e^{-|t-1|}$. Observe that $e^{-|t-1|}$ is $e^{-|t|}$ delayed by 1 second. For the last case you need to consider both the amplitude and the phase spectra.

4.2-3 Repeat Prob. 4.2-1 [parts (**a**), (**b**), and (**c**) only] for $m(t) = e^{-|t|}$ if the carrier is $\cos(10{,}000t - \pi/4)$. *Hint*: Use Eq. (3.36).

4.2-4 You are asked to design a DSB-SC modulator to generate a modulated signal $km(t) \cos \omega_c t$, where $m(t)$ is a signal band-limited to B Hz. Figure P4.2-4 shows a DSB-SC modulator available in the stock room. The carrier generator available generates not $\cos \omega_c t$, but $\cos^3 \omega_c t$. Explain whether you would be able to generate the desired signal using only this equipment. You may use any kind of filter you like.

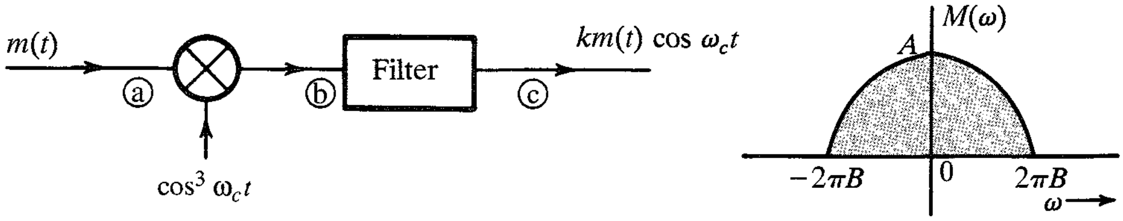
(a) What kind of filter is required in Fig. P4.2-4?

(b) Determine the signal spectra at points b and c, and indicate the frequency bands occupied by these spectra.

(c) What is the minimum usable value of ω_c?

(d) Would this scheme work if the carrier generator output were $\cos^2 \omega_c t$? Explain.

(e) Would this scheme work if the carrier generator output were $\cos^n \omega_c t$ for any integer $n \geq 2$?



(a) (b)

Figure P4.2-4

4.2-5 You are asked to design a DSB-SC modulator to generate a modulated signal $km(t) \cos \omega_c t$ with the carrier frequency $f_c = 300$ kHz ($\omega_c = 2\pi \times 300{,}000$). The following equipment is available in the stock room: (**i**) a signal generator of frequency 100 kHz; (**ii**) a ring modulator; (**iii**) a bandpass filter tuned to 300 kHz.

(a) Show how you can generate the desired signal.

(b) If the output of the modulator is $km(t) \cos \omega_c t$, find k.

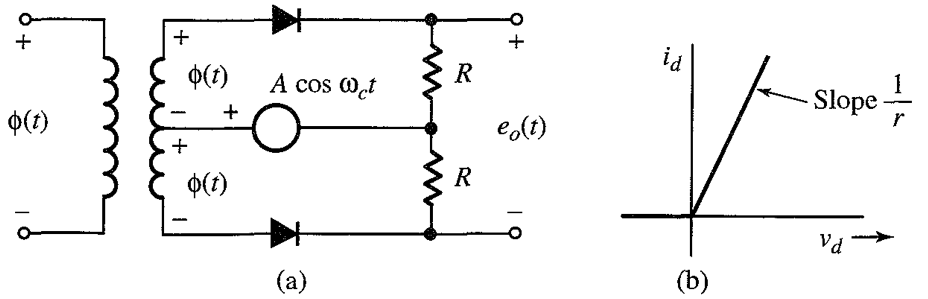


Figure P4.2-6

4.2-6 In Fig. P4.2-6, the input $\phi(t) = m(t)$, and the amplitude $A \gg |\phi(t)|$. The two diodes are identical with a resistance r ohms in the conducting mode and infinite resistance in the cutoff mode. Show that the output $e_o(t)$ is given by

$$e_o(t) = \frac{2R}{R+r}\, w(t)\, m(t)$$

where $w(t)$ is the switching periodic signal shown in Fig. 2.22a with period $2\pi/W_c$ seconds.

(a) Hence, show that this circuit can be used as a DSB-SC modulator.

(b) How would you use this circuit as a synchronous demodulator for DSB-SC signals.

4.2-7 In Fig. P4.2-6, if $\phi(t) = \sin(\omega_c t + \theta)$, and the output $e_0(t)$ is passed through a low-pass filter, then show that this circuit can be used as a phase detector, that is, a circuit that measures the phase difference between two sinusoids of the same frequency (ω_c). *Hint*: show that the filter output is a dc signal proportional to $\sin\theta$.

4.2-8 Two signals $m_1(t)$ and $m_2(t)$, both band-limited to 5000 rad/s, are to be transmitted simultaneously over a channel by the multiplexing scheme shown in Fig. P4.2-8. The signal at point b is the multiplexed signal, which now modulates a carrier of frequency 20,000 rad/s. The modulated signal at point c is transmitted over a channel.

(a) Sketch signal spectra at points a, b, and c.

(b) What must be the bandwidth of the channel?

(c) Design a receiver to recover signals $m_1(t)$ and $m_2(t)$ from the modulated signal at point c.

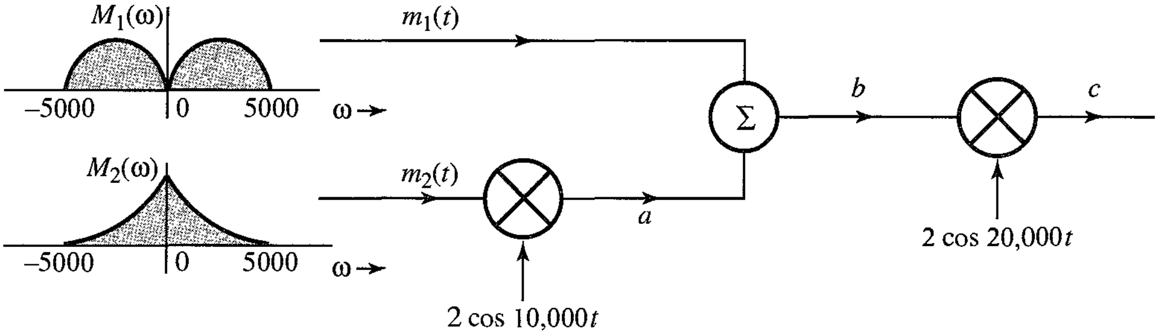


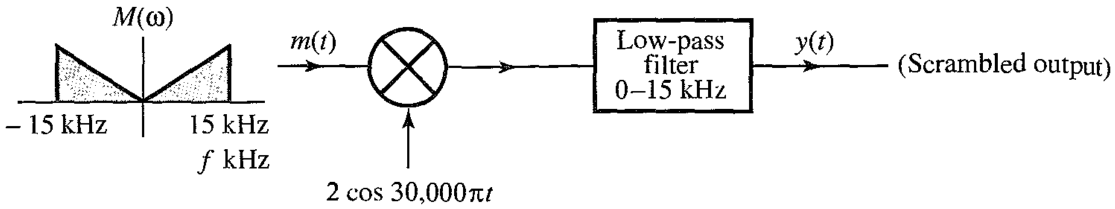
Figure P4.2-8

4.2-9 System shown in Fig. P4.2-9 is used for scrambling audio signals. The output $y(t)$ is the scrambled version of the input $m(t)$.

(a) Find the spectrum of the scrambled signal $y(t)$.

(b) Suggest a method of descrambling $y(t)$ to obtain $m(t)$.

A slightly modified version of this scrambler was first used commercially on the 25-mile radiotelephone circuit connecting Los Angeles and Santa Catalina island.

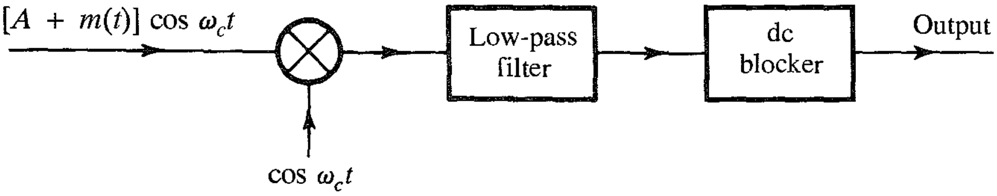


$M(\omega)$
-15 kHz
15 kHz
f kHz
$m(t)$
$2\cos 30{,}000\pi t$
Low-pass filter 0–15 kHz
$y(t)$
(Scrambled output)

Figure P4.2-9

4.2-10 A DSB-SC signal is given by $m(t)\cos(2\pi)10^6 t$. The carrier frequency of this signal, 1 MHz, is to be changed to 400 kHz. The only equipment available is one ring modulator, a bandpass filter centered at the frequency of 400 kHz, and one sine wave generator whose frequency can be varied from 150 to 210 kHz. Show how you can obtain the desired signal $cm(t)\cos(2\pi \times 400 \times 10^6 t)$ from $m(t)\cos(2\pi)10^6 t$. Determine the value of c.

4.3-1 Figure P4.3-1 shows a scheme for coherent (synchronous) demodulation. Show that this scheme can demodulate the AM signal $[A + m(t)]\cos \omega_c t$ regardless of the value of A.



$[A + m(t)]\cos \omega_c t$
Low-pass filter
dc blocker
Output
$\cos \omega_c t$

Figure P4.3-1

4.3-2 Sketch the AM signal $[A+m(t)]\cos \omega_c t$ for the periodic triangle signal $m(t)$ shown in Fig. P4.3-2 corresponding to the modulation index: **(a)** $\mu = 0.5$; **(b)** $\mu = 1$; **(c)** $\mu = 2$; **(d)** $\mu = \infty$. How do you interpret the case $\mu = \infty$?

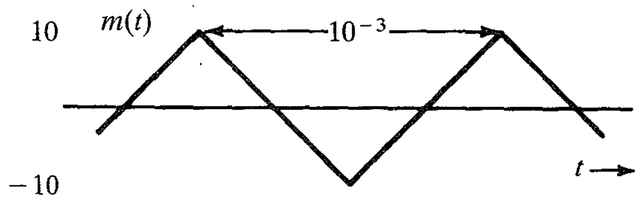


Figure P4.3-2

4.3-3 For the AM signal in Prob. 4.3-2 with $\mu = 0.8$:

(a) Find the amplitude and power of the carrier.

(b) Find the sideband power and the power efficiency η.

4.3-4 (a) Sketch the DSB-SC signal corresponding to $m(t) = \cos 2\pi t$.

(b) This DSB-SC signal $m(t)\cos\omega_c t$ is applied at the input of an envelope detector. Show that the output of the envelope detector is not $m(t)$, but $|m(t)|$. Show that, in general, if an AM signal $[A+m(t)]\cos\omega_c t$ is envelope-detected, the output is $|A+m(t)|$. Hence, show that the condition for recovering $m(t)$ from the envelope detector is $A+m(t)>0$ for all t.

4.3-5 Show that any scheme that can be used to generate DSB-SC can also generate AM. Is the converse true? Explain.

4.3-6 Show that any scheme that can be used to demodulate DSB-SC can also demodulate AM. Is the converse true? Explain.

4.3-7 In the text, the power efficiency of AM for a sinusoidal $m(t)$ was found. Carry out a similar analysis when $m(t)$ is a random binary signal as shown in Fig. P4.3-7 and $\mu = 1$. Sketch the AM signal with $\mu = 1$. Find the sideband's power and the total power (power of the AM signal) as well as their ratio (the power efficiency η).

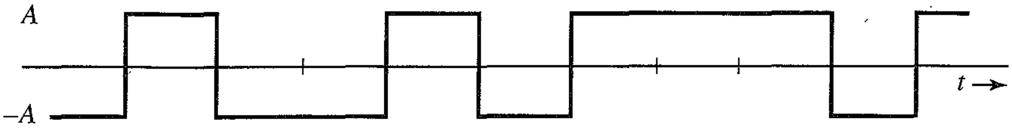


Figure P4.3-7

4.3-8 In the early days of radio, AM signals were demodulated by a crystal detector followed by a low-pass filter and a dc blocker, as shown in Fig. P4.3-8. Assume a crystal detector to be basically a squaring device. Determine the signals at points a, b, c, and d. Point out the distortion term in the output $y(t)$. Show that if $A \gg |m(t)|$, the distortion is small.

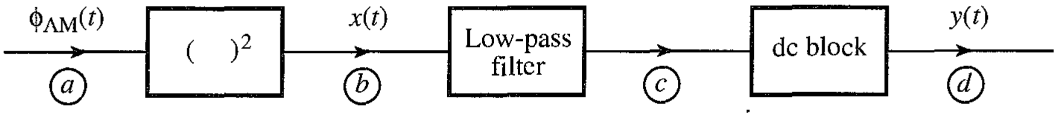


Figure P4.3-8

4.4-1 In a QAM system (Fig. 4.14), the locally generated carrier has a frequency error $\Delta\omega$ and a phase error δ; that is, the receiver carrier is $\cos[(\omega_c+\Delta\omega)t+\delta]$ or $\sin[(\omega_c+\Delta\omega)t+\delta]$. Show that the output of the upper receiver branch is

$$m_1(t)\cos[(\Delta\omega)t+\delta] - m_2(t)\sin[(\Delta\omega)t+\delta]$$

instead of $m_1(t)$, and the output of the lower receiver branch is

$$m_1(t)\sin[(\Delta\omega)t+\delta] + m_2(t)\cos[(\Delta\omega)t+\delta]$$

instead of $m_2(t)$.

4.5-1 A modulating signal $m(t)$ is given by:

(a) $m(t) = \cos 100t$

(b) $m(t) = \cos 100t + 2\cos 300t$

(c) $m(t) = \cos 100t \cos 500t$

In each case:

(i) Sketch the spectrum of $m(t)$.

(ii) Find and sketch the spectrum of the DSB-SC signal $2m(t)\cos 1000t$.

(iii) From the spectrum obtained in **(ii)**, suppress the LSB spectrum to obtain the USB spectrum.

(iv) Knowing the USB spectrum in **(ii)**, write the expression $\varphi_{\text{USB}}(t)$ for the USB signal.

(v) Repeat **(iii)** and **(iv)** to obtain the LSB signal $\varphi_{\text{LSB}}(t)$.

4.5-2 For the signals in Prob. 4.5-1, determine $\varphi_{\text{LSB}}(t)$ and $\varphi_{\text{USB}}(t)$ using Eq. (4.17) if the carrier frequency $\omega_c = 1000$. *Hint*: If $m(t)$ is a sinusoid, its Hilbert transform $m_h(t)$ is the sinusoid $m(t)$ phase-delayed by $\pi/2$ rad.

4.5-3 Find $\varphi_{\text{LSB}}(t)$ and $\varphi_{\text{USB}}(t)$ for the modulating signal $m(t) = B\,\text{sinc}\,(2\pi Bt)$ with $B = 1000$ and carrier frequency $\omega_c = 10,000\pi$. Follow these do-it-yourself steps:

(a) Sketch spectra of $m(t)$ and the corresponding DSB-SC signal $2m(t)\cos\omega_c t$.

(b) To find the LSB spectrum, suppress the USB in the DSB-SC spectrum found in **(a)**.

(c) Find the LSB signal $\varphi_{\text{LSB}}(t)$, which is the inverse Fourier transform of the LSB spectrum found in part **(b)**. Follow a similar procedure to find $\varphi_{\text{USB}}(t)$.

4.5-4 If $m_h(t)$ is the Hilbert transform of $m(t)$, then show that the Hilbert transform of $m_h(t)$ is $-m(t)$. (This shows that the inverse Hilbert transform operation is identical to the direct Hilbert transform operation with a negative sign.) Show also that the energies of $m(t)$ and $m_h(t)$ are identical. *Hint*: The Hilbert transform of $m(t)$ is obtained by passing $m(t)$ through a transfer function $H(\omega)$, whose amplitude and phase responses are shown in Fig. 4.17. The Hilbert transform of the Hilbert transform of $m(t)$ is obtained by passing $m(t)$ through $H(\omega)$ in cascade with $H(\omega)$.

4.5-5 An LSB signal is demodulated synchronously, as shown in Fig. P4.5-5. Unfortunately, the local carrier is not $2\cos\omega_c t$ as required, but is $2\cos[(\omega_c + \Delta\omega)t + \delta]$. Show that:

(a) When $\delta = 0$, the output $y(t)$ is the signal $m(t)$ with all its spectral components shifted (offset) by $\Delta\omega$. *Hint*: Observe that the output $y(t)$ is identical to the right-hand side of Eq. (4.17a) with ω_c replaced with $\Delta\omega$.

(b) When $\Delta\omega = 0$, the output is the signal $m(t)$ with phases of all its spectral components shifted by δ. *Hint*: Show that the output spectrum $Y(\omega) = M(\omega)e^{j\delta}$ for $\omega \geq 0$, and equal to $M(\omega)e^{-j\delta}$ when $\omega < 0$.

In each of these cases, explain the nature of distortion. *Hint*: For **(a)**, demodulation consists of shifting an LSB spectrum to the left and right by $\omega_c + \Delta\omega$, and low-pass filtering the result. For part **(b)**, use the expression (4.17b) for $\varphi_{\text{LSB}}(t)$ and multiply it by the local carrier $2\cos(\omega_c t + \delta)$, and low-pass filter the result.

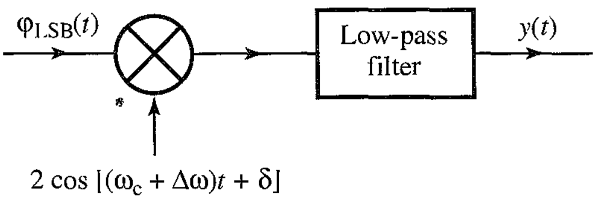

$\varphi_{\text{LSB}}(t)$
Low-pass filter
$y(t)$
$2\cos[(\omega_c + \Delta\omega)t + \delta]$

Figure P4.5-5

4.5-6 An USB signal is generated by using the phase-shift method (Fig. 4.20). If the input to this system is $m_h(t)$ instead of $m(t)$, what will be the output? Is this signal still an SSB signal with bandwidth equal to that of $m(t)$? Can this signal be demodulated [to get back $m(t)$]? If so, how?

4.6-1 A vestigial filter $H_i(\omega)$ shown in the transmitter of Fig. 4.22 has a transfer function as shown in Fig. P4.6-1. The carrier frequency is $f_c = 10$ kHz and the baseband signal bandwidth is 4 kHz. Find the corresponding transfer function of the equalizer filter $H_o(\omega)$ shown in the receiver of Fig. 4.22.

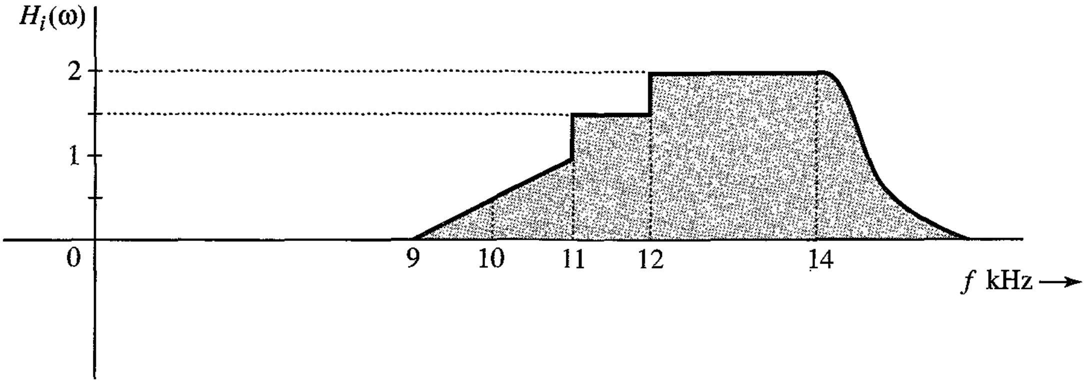


Figure P4.6-1

4.8-1 A transmitter transmits an AM signal with a carrier frequency of 1500 kHz. When an inexpensive radio receiver (which has a poor selectivity in its RF-stage bandpass filter) is tuned to 1500 kHz, the signal is heard loud and clear. This same signal is also heard (not as strong) at another dial setting. State, with reasons, at what frequency you will hear this station. The IF frequency is 455 kHz.

4.8-2 Consider a superheterodyne receiver designed to receive the frequency band of 1 to 30 MHz with IF frequency 8 MHz. What is the range of frequencies generated by the local oscillator for this receiver? An incoming signal with carrier frequency 10-MHz is received at the 10 MHz setting. At this setting of the receiver we also get interference from a signal with some other carrier frequency if the receiver RF stage bandpass filter has poor selectivity. What is the carrier frequency of the interfering signal?

5

ANGLE (EXPONENTIAL) MODULATION

In AM signals, the amplitude of a carrier is modulated by a signal $m(t)$, and, hence, the information content of $m(t)$ is in the amplitude variations of the carrier. Because a sinusoidal signal is described by amplitude and angle (which includes frequency and phase), there exists a possibility of carrying the same information by varying the angle of the carrier. This chapter explores such a possibility.

A Historical Note

In the twenties, broadcasting was in its infancy. However, there was a constant search for techniques that will reduce noise (static). Now, since the noise power is proportional to the modulated signal bandwidth (sidebands), the attempt was focused on finding a modulation scheme that will reduce the bandwidth. It was rumored that a new method had been discovered for eliminating sidebands (no sidebands, no bandwidth!). The idea of **frequency modulation (FM)**, where the carrier frequency would be varied in proportion to the message $m(t)$, appeared quite intriguing. The carrier frequency $\omega(t)$ would be varied with time so that $\omega(t) = \omega_c + km(t)$, where k is an arbitrary constant. If the peak amplitude of $m(t)$ is m_p, then the maximum and minimum values of the carrier frequency would be $\omega_c + km_p$ and $\omega_c - km_p$, respectively. Hence, the spectral components would remain within this band with a bandwidth $2km_p$ centered at ω_c. The bandwidth is controlled by the arbitrary constant k, whose value can be selected as we please. By using an arbitrarily small k, we could make the information bandwidth arbitrarily small. This was a passport to communication heaven. Unfortunately, the experimental results showed that something was seriously wrong somewhere. The FM bandwidth was found to be always greater than (at best equal to) the AM bandwidth. In some cases, its bandwidth was several times that of AM. Where is the fallacy in this reasoning? We shall soon find out.

5.1 CONCEPT OF INSTANTANEOUS FREQUENCY

By definition, a sinusoidal signal has a constant frequency, and, hence, the variation of frequency with time appears to be contradictory to the conventional definition of a sinusoidal signal frequency. We must extend the concept of a sinusoid to a generalized function whose frequency may vary with time.

In FM we wish to vary the carrier frequency in proportion to the modulating signal $m(t)$. This means the carrier frequency is changing continuously every instant. Prima facie, this does not make much sense because to define a frequency, we must have a sinusoidal signal at least over one cycle (or a half-cycle or a quarter-cycle) with the same frequency. This problem reminds us of our first encounter with the concept of **instantaneous velocity** in our beginning mechanics course. Until that time, we were used to thinking of velocity as being constant over an interval, and we were incapable of even imagining that velocity could vary at each instant. But with some mental struggle, the idea gradually sinks in. We never forget, however, the wonder and amazement that was caused by the idea when it was first introduced. A similar experience awaits the reader with the concept of **instantaneous frequency**.

Let us consider a generalized sinusoidal signal $\varphi(t)$ given by

$$\varphi(t) = A\cos\theta(t) \tag{5.1}$$

where $\theta(t)$ is the **generalized angle** and is a function of t. Figure 5.1 shows a hypothetical case of $\theta(t)$. The generalized angle for a conventional sinusoid $A\cos(\omega_c t + \theta_0)$ is $\omega_c t + \theta_0$. This is a straight line with a slope ω_c and intercept θ_0, as shown in Fig. 5.1. The plot of $\theta(t)$ for the hypothetical case happens to be tangential to the angle $(\omega_c t + \theta_0)$ at some instant t. The crucial point is that over a small interval $\Delta t \to 0$, the signal $\varphi(t) = A\cos\theta(t)$ and the sinusoid $A\cos(\omega_c t + \theta_0)$ are identical; that is,

$$\varphi(t) = A\cos(\omega_c t + \theta_0) \qquad t_1 < t < t_2$$

We are certainly justified in saying that over this small interval Δt, the frequency of $\varphi(t)$ is ω_c. Because $(\omega_c t + \theta_0)$ is tangential to $\theta(t)$, the frequency of $\varphi(t)$ is the slope of its angle $\theta(t)$ over this small interval. We can generalize this concept at every instant and say that the instantaneous frequency ω_i at any instant t is the slope of $\theta(t)$ at t. Thus, for $\varphi(t)$ in Eq. (5.1),

$$\omega_i(t) = \frac{d\theta}{dt} \tag{5.2a}$$

$$\theta(t) = \int_{-\infty}^{t} \omega_i(\alpha)\, d\alpha \tag{5.2b}$$

Now we can see the possibility of transmitting the information of $m(t)$ by varying the angle θ of a carrier. Such techniques of modulation, where the angle of the carrier is varied in some manner with a modulating signal $m(t)$, are known as **angle modulation** or **exponential modulation**. Two simple possibilities are: **phase modulation (PM)** and **frequency modulation (FM)**. In PM, the angle $\theta(t)$ is varied linearly with $m(t)$:

$$\theta(t) = \omega_c t + \theta_0 + k_p m(t)$$

where k_p is a constant and ω_c is the carrier frequency. Assuming $\theta_0 = 0$, without loss of generality,

$$\theta(t) = \omega_c t + k_p m(t) \tag{5.3a}$$

The resulting PM wave is

$$\varphi_{\text{PM}}(t) = A\cos[\omega_c t + k_p m(t)] \tag{5.3b}$$

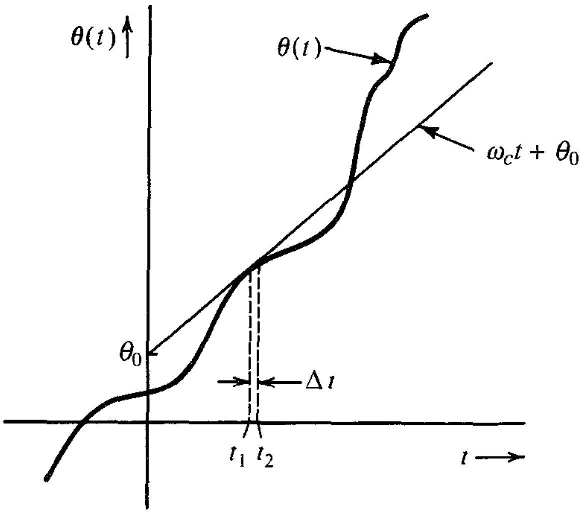


Figure 5.1 Concept of instantaneous frequency.

The instantaneous frequency $\omega_i(t)$ in this case is given by

$$\omega_i(t) = \frac{d\theta}{dt} = \omega_c + k_p \dot{m}(t) \tag{5.3c}$$

Hence, in PM, the instantaneous frequency ω_i varies linearly with the derivative of the modulating signal. If the instantaneous frequency ω_i is varied linearly with the modulating signal, we have FM. Thus, in FM the instantaneous frequency ω_i is

$$\omega_i(t) = \omega_c + k_f m(t) \tag{5.4a}$$

where k_f is a constant. The angle $\theta(t)$ is now

$$\begin{aligned}\theta(t) &= \int_{-\infty}^{t} [\omega_c + k_f m(\alpha)]\, d\alpha \\ &= \omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha \end{aligned} \tag{5.4b}$$

Here we have assumed the constant term in $\theta(t)$ to be zero without loss of generality. The FM wave is

$$\varphi_{\text{FM}}(t) = A \cos \left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha\right] \tag{5.4c}$$

Generalized Concept of Angle Modulation

From Eqs. (5.3b) and (5.4c), it is apparent that PM and FM are not only very similar but are inseparable. Replacing $m(t)$ in Eq. (5.3b) with $\int m(t)$ changes PM into FM. Thus, a signal that is an FM wave corresponding to $m(t)$ is also the PM wave corresponding to $\int m(\alpha)\, d\alpha$ (Fig. 5.2a). Similarly, a PM wave corresponding to $m(t)$ is the FM wave corresponding to $\dot{m}(t)$ (Fig. 5.2b). Therefore, by looking at an angle-modulated carrier, there is no way of telling whether it is FM or PM. In fact, it is meaningless to ask an angle-modulated wave whether it is FM or PM. An analogous practical situation would be to ask a person (who is married, with children) whether he is a father or a son. The person would be puzzled because he is both, a father (of his child) and a son (of his father).

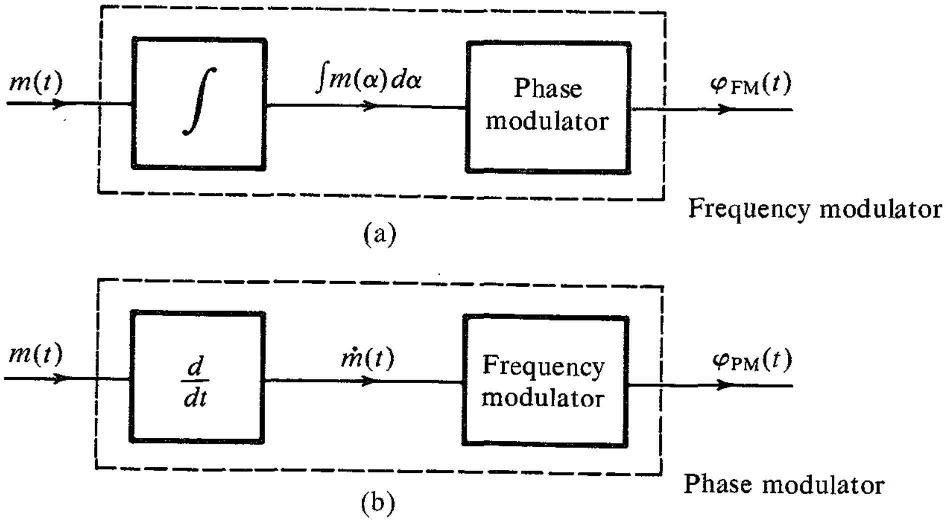


Figure 5.2 Phase and frequency modulation are inseparable.

Equations (5.3b) and (5.4c) show that in both PM and FM the angle of a carrier is varied in proportion to some measure of $m(t)$. In PM, it is directly proportional to $m(t)$, whereas in FM, it is proportional to the integral of $m(t)$. But why should we limit ourselves only to these cases? We have an infinite number of possible ways of generating a measure of $m(t)$. If we restrict the choice to a linear operator, then a measure of $m(t)$ can be obtained as the output of a suitable linear (time-invariant) system with $m(t)$ as its input, as shown in Fig. 5.3. The system transfer function is $H(s)$ and its impulse response is $h(t)$. The output of this system, $\psi(t)$, is a measure of $m(t)$. This is a reversible operation; that is, $m(t)$ can be recovered from $\psi(t)$ by passing it through a system of the transfer function $1/H(s)$.

The generalized angle-modulated carrier $\varphi_{\text{EM}}(t)$ can be expressed as

$$\varphi_{\text{EM}}(t) = A\cos[\omega_c t + \psi(t)] \tag{5.5a}$$

$$= A\cos\left[\omega_c t + \int_{-\infty}^{t} m(\alpha)\,h(t-\alpha)\,d\alpha\right] \tag{5.5b}$$

If $h(t) = k_p\delta(t)$, this equation reduces to Eq. (5.3b), and we have the conventional PM. Similarly, if $h(t) = k_f u(t)$, the equation reduces to Eq. (5.4c), resulting in conventional FM. Now, FM and PM are just two possibilities (out of an infinite number.) We shall see later that the optimum performance system is neither FM nor PM, but something else, depending on the modulating signal spectrum and the channel characteristics.

The generalized angle modulation concept is useful because it shows the convertibility of one type of angle modulation (such as PM) to another (such as FM). This is quite clear from Fig. 5.2. For instance, we show later that the bandwidth of FM is approximately $2k_f m_p$, where m_p is the peak amplitude of $m(t)$. We can derive the equivalent result for PM by referring to Fig. 5.2b, which shows that PM is actually the FM when the modulating signal is $\dot{m}(t)$. Clearly, the bandwidth of PM is approximately $2k_p m'_p$, where m'_p is the peak amplitude of $\dot{m}(t)$. This

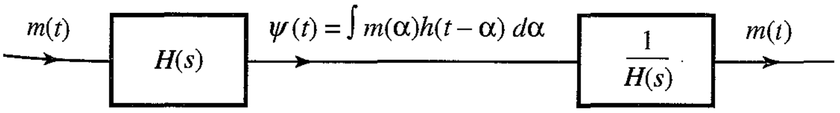


Figure 5.3 Generalized exponential modulation.

shows that if we analyze one type of angle modulation (such as FM), we can readily extend those results to any other kind. Historically, the angle modulation concept began with FM, and in this chapter we shall primarily analyze FM, with occasional discussion of PM. But this does not mean that FM is superior to other kinds of angle modulation. On the contrary, for most practical signals, PM is superior to FM. Actually, the optimum performance is realized neither by PM nor by FM, but by something in between.

This discussion also shows that we need not discuss methods of generation and demodulation of each type of modulation. From Fig. 5.2, it is clear that PM can be generated by an FM generator, and FM can be generated by a PM generator. One of the methods of generating FM in practice (the Armstrong indirect-FM system) actually integrates $m(t)$ and uses it to phase-modulate a carrier (see Fig. 5.6).

EXAMPLE 5.1 Sketch FM and PM waves for the modulating signal $m(t)$ shown in Fig. 5.4a. The constants k_f and k_p are $2\pi \times 10^5$ and 10π, respectively, and the carrier frequency f_c is 100 MHz.

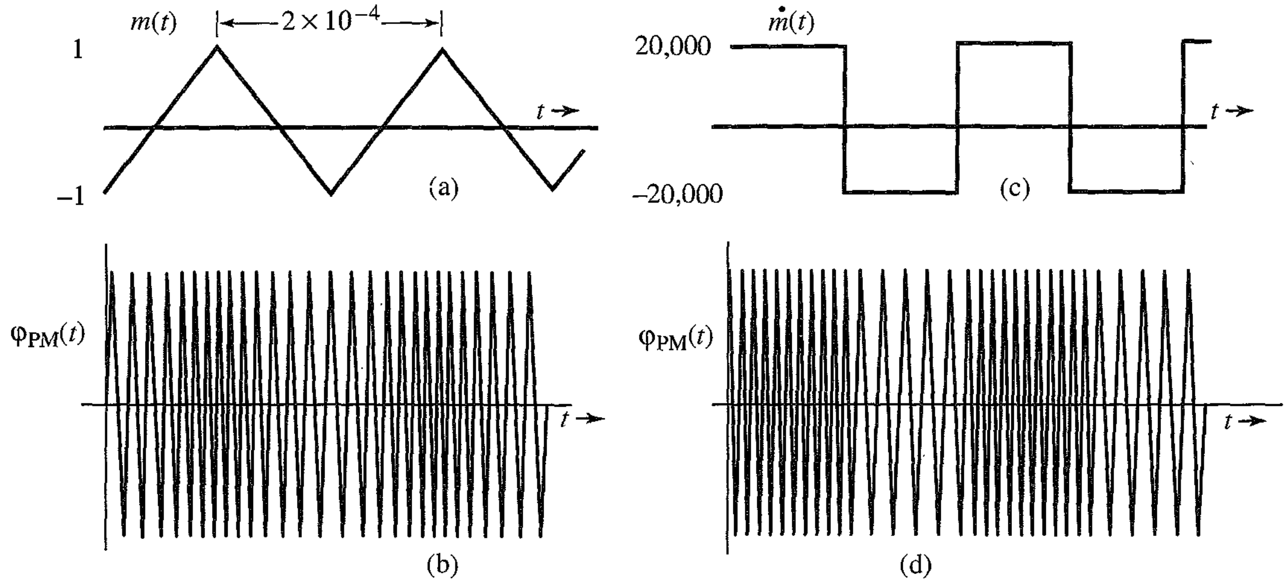


Figure 5.4 FM and PM waveforms.

For FM:

$$\omega_i = \omega_c + k_f m(t)$$

Dividing throughout by 2π, we have the equation in terms of the variable f (frequency in hertz). The instantaneous frequency f_i is

$$\begin{aligned} f_i &= f_c + \frac{k_f}{2\pi} m(t) \\ &= 10^8 + 10^5 m(t) \\ (f_i)_{\text{min}} &= 10^8 + 10^5\,[m(t)]_{\text{min}} = 99.9 \text{ MHz} \\ (f_i)_{\text{max}} &= 10^8 + 10^5\,[m(t)]_{\text{max}} = 100.1 \text{ MHz} \end{aligned}$$

Because $m(t)$ increases and decreases linearly with time, the instantaneous frequency increases linearly from 99.9 to 100.1 MHz over a half-cycle and decreases linearly from 100.1 to 99.9 MHz over the remaining half-cycle of the modulating signal (Fig. 5.4b).

For PM: PM for $m(t)$ is FM for $\dot{m}(t)$. This also follows from Eq. (5.3c).

$$\begin{aligned} f_i &= f_c + \frac{k_p}{2\pi}\dot{m}(t) \\ &= 10^8 + 5\,\dot{m}(t) \\ (f_i)_{\min} &= 10^8 + 5\,[\dot{m}(t)]_{\min} = 10^8 - 10^5 = 99.9 \text{ MHz} \\ (f_i)_{\max} &= 10^8 + 5\,[\dot{m}(t)]_{\max} = 100.1 \text{ MHz} \end{aligned}$$

Because $\dot{m}(t)$ switches back and forth from a value of $-20{,}000$ to $20{,}000$, the carrier frequency switches back and forth from 99.9 to 100.1 MHz every half-cycle of $\dot{m}(t)$, as shown in Fig. 5.4d.

This indirect method of sketching PM [using $\dot{m}(t)$ to frequency-modulate a carrier] works as long as $m(t)$ is a continuous signal. If $m(t)$ is discontinuous, $\dot{m}(t)$ contains impulses, and this method fails. In such a case, a direct approach should be used. This is demonstrated in the next example.

EXAMPLE 5.2 Sketch FM and PM waves for the digital modulating signal $m(t)$ shown in Fig. 5.5a. The constants k_f and k_p are $2\pi \times 10^5$ and $\pi/2$, respectively, and $f_c = 100$ MHz.

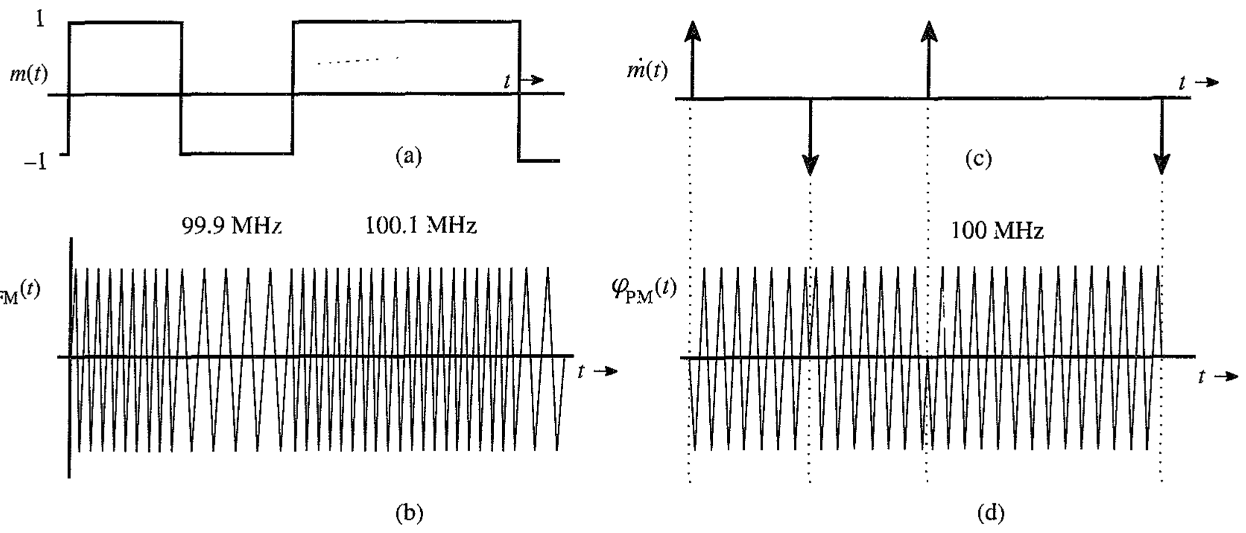


Figure 5.5 FM and PM waveforms.

For FM:

$$f_i = f_c + \frac{k_f}{2\pi}m(t) = 10^8 + 10^5 m(t)$$

Because $m(t)$ switches from 1 to -1 and vice versa, the FM wave frequency switches back and forth between 99.9 MHz and 100.1 MHz, as shown in Fig. 5.5b. This scheme of carrier frequency modulation by a digital signal (Fig. 5.5b) is called **frequency-shift keying (FSK)** because information digits are transmitted by shifting the carrier frequency (see Sec. 7.8).

For PM:

$$f_i = f_c + \frac{k_p}{2\pi}\dot{m}(t) = 10^8 + \frac{1}{4}\dot{m}(t)$$

The derivative $\dot{m}(t)$ (Fig. 5.5c) contains impulses of strength ± 2, and it is not immediately apparent how an instantaneous frequency can be changed by an infinite amount and then changed back to the original frequency in zero time. Let us consider the direct approach:

$$\begin{aligned}\varphi_{\text{PM}}(t) &= A \cos\left[\omega_c t + k_p m(t)\right] \\ &= A \cos\left[\omega_c t + \frac{\pi}{2} m(t)\right] \\ &= \begin{cases} A \sin \omega_c t & \text{when } m(t) = -1 \\ -A \sin \omega_c t & \text{when } m(t) = 1 \end{cases}\end{aligned}$$

This PM wave is shown in Fig. 5.5d. This scheme of carrier PM by a digital signal is called **phase-shift keying (PSK)** because information digits are transmitted by shifting the carrier phase. Note that PSK may also be viewed as a DSB-SC modulation by $m(t)$.

The PM wave $\varphi_{\text{PM}}(t)$ in this case has phase discontinuities at instants where impulses of $\dot{m}(t)$ are located. At these instants, the carrier phase shifts by π instantaneously. A finite phase shift in zero time implies infinite instantaneous frequency at these instants. This agrees with our observation about $\dot{m}(t)$.

The amount of phase discontinuity in $\varphi_{\text{PM}}(t)$ at the instant where $m(t)$ is discontinuous is $k_p m_d$, where m_d is the amount of discontinuity in $m(t)$ at that instant. In the present example, the amplitude of $m(t)$ changes by 2 (from -1 to 1) at the discontinuity. Hence, the phase discontinuity in $\varphi_{\text{PM}}(t)$ is $k_p m_d = (\pi/2) \times 2 = \pi$ rad, which confirms our earlier result.

When $m(t)$ is a digital signal (as in Fig. 5.5a), $\varphi_{\text{PM}}(t)$ shows a phase discontinuity where $m(t)$ has a jump discontinuity. We shall now show that in such a case the phase deviation $k_p m(t)$ must be restricted to a range $(-\pi, \pi)$ in order to avoid ambiguity in demodulation. For example, if k_p were $3\pi/2$ in the present example, then

$$\varphi_{\text{PM}}(t) = A \cos\left[\omega_c t + \frac{3\pi}{2} m(t)\right]$$

In this case $\varphi_{\text{PM}}(t) = A \sin \omega_c t$ when $m(t) = 1$ or $-1/3$. This will certainly cause ambiguity at the receiver when $A \sin \omega_c t$ is received. Such ambiguity never arises if $k_p m(t)$ is restricted to the range $(-\pi, \pi)$.

What causes this ambiguity? When $m(t)$ has jump discontinuities, the phase of $\varphi_{\text{PM}}(t)$ changes instantaneously. Because a phase $\varphi_o + 2n\pi$ is indistinguishable from the phase φ_o, ambiguities will be inherent in the demodulator unless the phase variations are limited to the range $(-\pi, \pi)$. This means k_p should be small enough to restrict the phase change $k_p m(t)$ to the range $(-\pi, \pi)$.

No such restriction on k_p is required if $m(t)$ is continuous. In this case the phase change is not instantaneous, but gradual over a time, and a phase $\varphi_o + 2n\pi$ will exhibit n additional carrier cycles over the case of phase of only φ_o. We can detect the PM wave by using an FM demodulator followed by an integrator (see Prob. 5.4-1). The additional n cycles will be detected by the FM demodulator, and the subsequent integration will yield a phase $2n\pi$. Hence, the phases φ_o and $\varphi_o + 2n\pi$ can be detected without ambiguity. This conclusion can also be verified from Example 5.1, where the maximum phase change $\Delta\varphi = 10\pi$.

Because a band-limited signal cannot have jump discontinuities, we can say that when $m(t)$ is band-limited, k_p has no restrictions.

Power of an Angle-Modulated Wave

Although the instantaneous frequency and phase of an angle-modulated wave can vary with time, the amplitude A always remains constant. Hence, the power of an angle-modulated wave (PM or FM) is always $A^2/2$, regardless of the value of k_p or k_f.

5.2 BANDWIDTH OF ANGLE-MODULATED WAVES

In order to determine the bandwidth of an FM wave, let us define

$$a(t) = \int_{-\infty}^{t} m(\alpha)\, d\alpha \tag{5.6}$$

and

$$\hat{\varphi}_{\text{FM}}(t) = A\, e^{j[\omega_c t + k_f a(t)]} = A e^{jk_f a(t)} e^{j\omega_c t} \tag{5.7a}$$

Now

$$\varphi_{\text{FM}}(t) = \text{Re}\, \hat{\varphi}_{\text{FM}}(t) \tag{5.7b}$$

Expanding the exponential $e^{jk_f a(t)}$ in Eq. (5.7a) in power series yields

$$\hat{\varphi}_{\text{FM}}(t) = A\left[1 + jk_f a(t) - \frac{k_f^2}{2!}a^2(t) + \cdots + j^n \frac{k_f^n}{n!}a^n(t) + \cdots\right] e^{j\omega_c t} \tag{5.8a}$$

and

$$\begin{aligned}\varphi_{\text{FM}}(t) &= \text{Re}\,[\hat{\varphi}_{\text{FM}}(t)] \\ &= A\left[\cos\omega_c t - k_f a(t)\sin\omega_c t - \frac{k_f^2}{2!}a^2(t)\cos\omega_c t + \frac{k_f^3}{3!}a^3(t)\sin\omega_c t + \cdots\right]\end{aligned} \tag{5.8b}$$

The modulated wave consists of an unmodulated carrier plus various amplitude-modulated terms, such as $a(t)\sin\omega_c t$, $a^2(t)\cos\omega_c t$, $a^3(t)\sin\omega_c t$, ... The signal $a(t)$ is an integral of $m(t)$. If $M(\omega)$ is band-limited to B, $A(\omega)$ is also band-limited* to B. The spectrum of $a^2(t)$

---

* This is because integration is a linear operation equivalent to passing a signal through a transfer function $1/j\omega$. Hence, if $M(\omega)$ is band-limited to B, $A(\omega)$ must also be band-limited to B.

is simply $A(\omega) * A(\omega)/2\pi$ and is band-limited to $2B$. Similarly, the spectrum of $a^n(t)$ is band-limited to nB. Hence, the spectrum consists of an unmodulated carrier plus spectra of $a(t)$, $a^2(t)$, $\dots$, $a^n(t)$, $\dots$, centered at ω_c. Clearly, the modulated wave is not band-limited. It has an infinite bandwidth and is not related to the modulating-signal spectrum in any simple way, as was the case in AM.

Although the theoretical bandwidth of an FM wave is infinite, we shall see that most of the modulated-signal power resides in a finite bandwidth. There are two distinct possibilities in terms of bandwidths—narrow-band FM and wide-band FM.

Narrow-Band Angle Modulation

Unlike AM, angle modulation is nonlinear. The principle of superposition does not apply. This may be verified from the fact that

$$A\cos\{\omega_c t + k_f[a_1(t) + a_2(t)]\} \neq A\cos[\omega_c t + k_f a_1(t)] + A\cos[\omega_c t + k_f a_2(t)]$$

The principle of superposition does not hold. If, however, k_f is very small (that is, if $|k_f a(t)| \ll 1$), then all but the first two terms in Eq. (5.8) are negligible, and we have

$$\varphi_{\mathrm{FM}}(t) \simeq A[\cos\omega_c t - k_f a(t)\sin\omega_c t] \tag{5.9}$$

This is a linear modulation. This expression is similar to that of the AM wave. Because the bandwidth of $a(t)$ is B, the bandwidth of $\varphi_{\mathrm{FM}}(t)$ in Eq. (5.9) is only $2B$. For this reason, the case ($|k_f a(t)| \ll 1$) is called **narrow-band FM (NBFM)**. The **narrow-band PM (NBPM)** case is similarly given by

$$\varphi_{\mathrm{PM}}(t) \simeq A[\cos\omega_c t - k_p m(t)\sin\omega_c t] \tag{5.10}$$

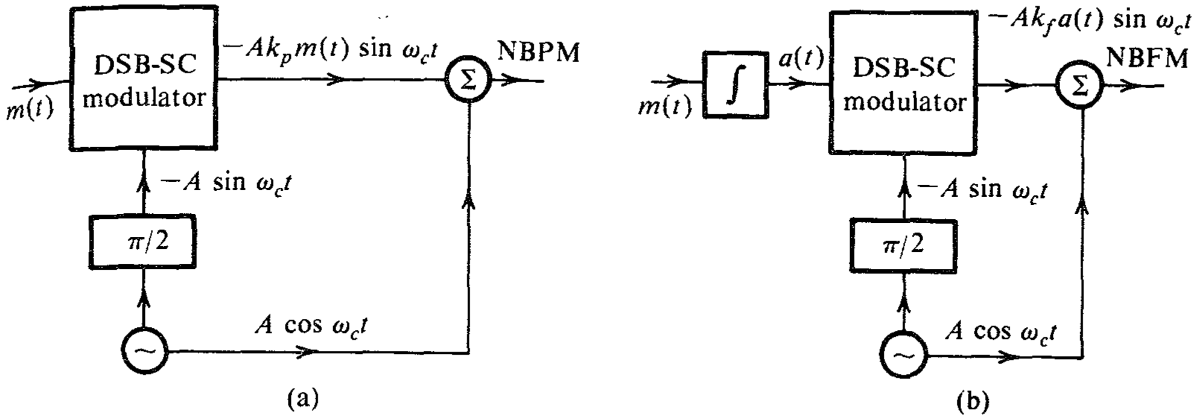
A comparison of NBFM [Eq. (5.9)] with AM [Eq. (4.8a)] brings out clearly the similarities and differences between the two types of modulation. Both cases have a carrier term and sidebands centered at $\pm\,\omega_c$. The modulated-signal bandwidths are identical (viz., $2B$). The sideband spectrum for FM has a phase shift of $\pi/2$ with respect to the carrier, whereas that of AM is in phase with the carrier. It must be remembered, however, that despite apparent similarities, the AM and FM signals have very different waveforms. In an AM signal, the frequency is constant and the amplitude varies with time, whereas in an FM signal, the amplitude is constant and the frequency varies with time.

Equations (5.9) and (5.10) suggest a possible method of generating narrow-band FM and PM signals by using DSB-SC modulators. The block-diagram representation of such systems is shown in Fig. 5.6.

Wide-Band FM (WBFM): The Fallacy Exposed

If the deviation in the carrier frequency is large enough [i.e., if the constant k_f is chosen large enough so that the condition $|k_f a(t)| \ll 1$ is not satisfied], we cannot ignore the higher order terms in Eq. (5.8b), and the preceding analysis becomes too complicated to lead to a fruitful solution. We shall take here the route of the pioneers, who by their intuitively simple reasoning came to grief in estimating the FM bandwidth. If we could discover the fallacy in their reasoning, we would have a chance of obtaining a better estimate of the wide-band FM bandwidth.

Consider an $m(t)$ that is band-limited to B Hz. This signal is approximated by a staircase signal $\hat{m}(t)$, as shown in Fig. 5.7a. The signal $m(t)$ is now approximated by pulses of constant



(a) (b)

Figure 5.6 Narrow-band PM and FM wave generation.

amplitudes. For convenience, each of these pulses will be called a "cell." It is relatively easy to analyze FM corresponding to $\hat{m}(t)$ because it has constant amplitudes. To ensure that $\hat{m}(t)$ has all the information of $m(t)$, the cell width in $\hat{m}(t)$ must be no greater than the Nyquist interval of $1/2B$ seconds. Thus, $m(t)$ is approximated by constant-amplitude pulses (cells) of width $T = 1/2B$ seconds. Consider a typical cell starting at $t = t_k$. This cell has a constant amplitude $m(t_k)$. Hence, the FM signal corresponding to this cell is a sinusoid of frequency $\omega_c + k_f m(t_k)$ and duration $T = 1/2B$, as shown in Fig. 5.7b. The FM signal for $\hat{m}(t)$ consists of a sequence of such sinusoidal pulses corresponding to various cells of $\hat{m}(t)$.

The FM spectrum for $\hat{m}(t)$ consists of the sum of the Fourier transforms of these sinusoidal pulses corresponding to all the cells. The Fourier transform of a sinusoidal pulse in Fig. 5.7b (corresponding to the kth cell) is a sinc function shown shaded in Fig. 5.7c (see Example 3.12, Fig. 3.22d with $T = 1/2B$). Note that the spectrum of this pulse is spread out on either side of its frequency $\omega_c + k_f m(t_k)$ by $2\pi/T = 4\pi B$. Figure 5.7c shows the spectra of sinusoidal pulses corresponding to various cells. The minimum and the maximum amplitudes of the cells are $-m_p$ and m_p, respectively. Hence, the minimum and maximum frequencies of the sinusoidal pulses corresponding to the FM signal for all the cells are $\omega_c - k_f m_p$ and $\omega_c + k_f m_p$, respectively. Moreover, the spectrum for each sinusoid spreads out on either side of its frequency by $4\pi B$ rad/s, as shown in Fig. 5.7c. Hence, the maximum and the minimum significant frequencies in this spectrum are $\omega_c + k_f m_p + 4\pi B$ and $\omega_c - k_f m_p - 4\pi B$, respectively. The spectrum bandwidth is the difference $2k_f m_p + 8\pi B$.

We can now understand the fallacy in the reasoning of the pioneers. The maximum and minimum carrier frequencies are $\omega_c + k_f m_p$ and $\omega_c - k_f m_p$, respectively. Hence, it was reasoned that the spectral components must also lie in this range, resulting the FM bandwidth of $2k_f m_p$. The implicit assumption was that a sinusoid of frequency ω has its entire spectrum concentrated at ω. Unfortunately, this is true only of the everlasting sinusoid because the Fourier transform of such a sinusoid is an impulse at ω. For a sinusoid of finite duration T seconds, the spectrum is spread out on either side of ω by $2\pi/T$, as shown in Example 3.12. The pioneers had missed this spreading effect.

The deviation of the carrier frequency is $\pm k_f m_p$. We shall denote the carrier frequency deviation by $\Delta\omega$. Thus,

$$\Delta\omega = k_f m_p \tag{5.11}$$

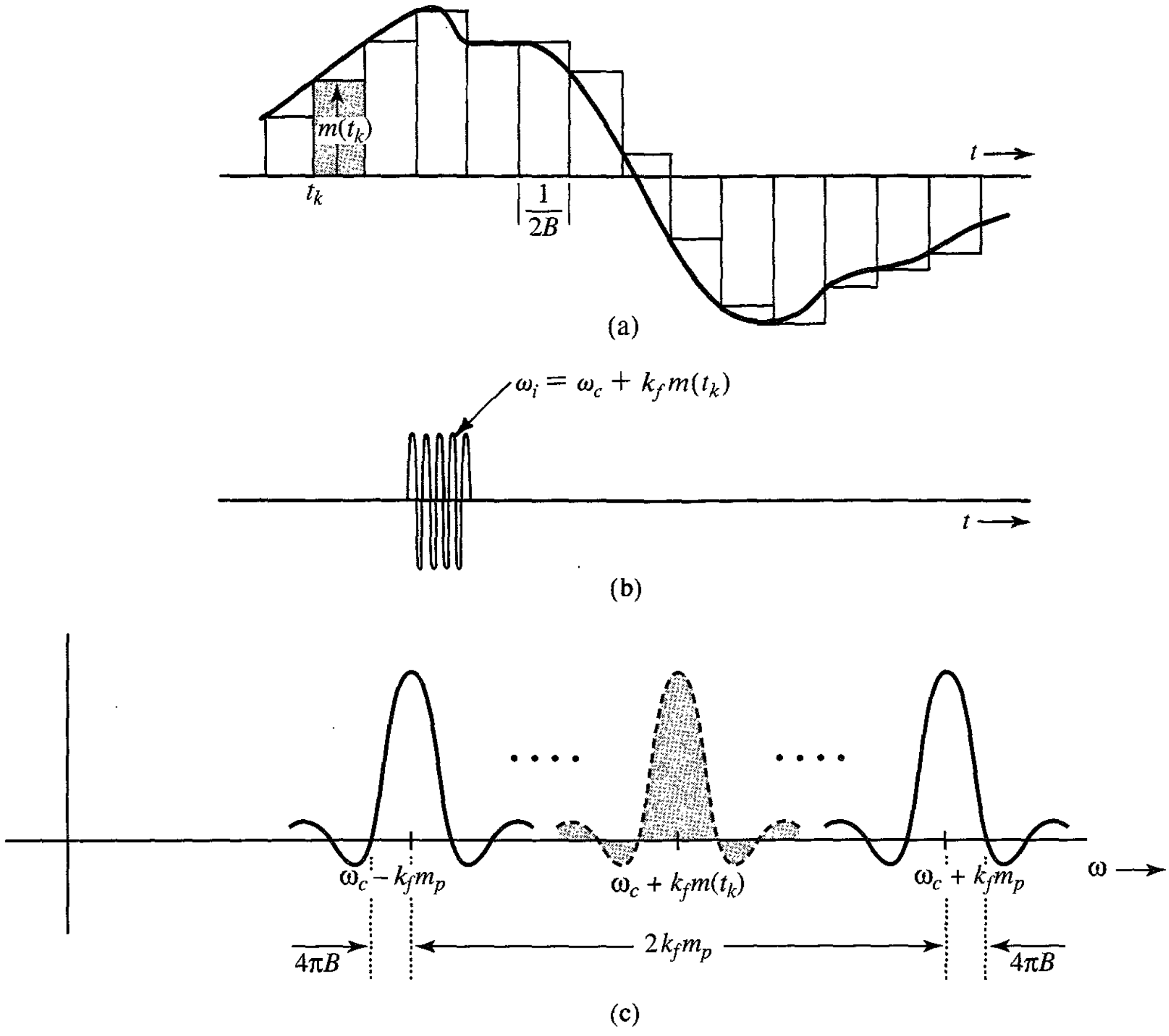


Figure 5.7 Estimation of FM wave bandwidth.

The carrier frequency deviation in hertz will be denoted by Δf. Thus,

$$\Delta f = \frac{k_f m_p}{2\pi}$$

The estimated FM bandwidth (in hertz) can be expressed as

$$\begin{aligned} B_{\text{FM}} &= \frac{1}{2\pi}(2k_f m_p + 8\pi B) \\ &= 2(\Delta f + 2B) \end{aligned} \tag{5.12}$$

The bandwidth estimate thus obtained is somewhat higher than the actual value because this is the bandwidth corresponding to the staircase approximation of $m(t)$, not the actual $m(t)$, which is considerably smoother. Hence, the actual bandwidth is somewhat smaller than this value. Therefore, we must readjust our bandwidth estimation. In order to make this midcourse correction, we observe that for the narrow-band case, k_f is very small. Hence, Δf is very small (compared to B). In this case we can ignore the Δf term in Eq. (5.12) with the result

$$B_{\text{FM}} \approx 4B$$

But we have shown earlier that for narrow-band, the FM bandwidth is $2B$ Hz. This indicates that a better bandwidth estimate is

$$B_{\text{FM}} = 2(\Delta f + B) \tag{5.13a}$$

$$= 2\left(\frac{k_f m_p}{2\pi} + B\right) \tag{5.13b}$$

This is precisely the result obtained by Carson,[1] who investigated this problem rigorously for tone modulation [sinusoidal $m(t)$]. This formula goes under the name **Carson's rule** in the literature. Observe that for a truly wide-band case, where $\Delta f \gg B$, Eqs. (5.13) can be approximated as

$$B_{\text{FM}} \approx 2\Delta f \qquad \Delta f \gg B \tag{5.14}$$

Because $\Delta\omega = k_f m_p$, this formula is precisely what the pioneers had used for FM bandwidth. The only mistake was in thinking that this formula will hold for all cases, especially for the narrow-band case, where $\Delta f \ll B$.

We define a deviation ratio β as

$$\beta = \frac{\Delta f}{B} \tag{5.15}$$

Carson's rule can be expressed in terms of the deviation ratio as

$$B_{\text{FM}} = 2B(\beta + 1) \tag{5.16}$$

The deviation ratio controls the amount of modulation and, consequently, plays a role similar to the modulation index in AM. Indeed, for the special case of tone-modulated FM, the deviation ratio β is called the **modulation index.**

Phase Modulation

All the results derived for FM can be directly applied to PM. Thus, for PM, the instantaneous frequency is given by

$$\omega_i = \omega_c + k_p \dot{m}(t)$$

Therefore, the frequency deviation $\Delta\omega$ is given by

$$\Delta\omega = k_p m'_p \tag{5.17a}$$

where*

$$m'_p = [\dot{m}(t)]_{\max} \tag{5.17b}$$

---

* We are assuming that $|\dot{m}(t)_{\min}| = m'_p$.

Therefore,*

$$B_{\text{PM}} = 2(\Delta f + B) \tag{5.18a}$$

$$= 2\left(\frac{k_p m'_p}{2\pi} + B\right) \tag{5.18b}$$

One interesting aspect of FM is that $\Delta\omega = k_f m_p$ depends only on the peak value of $m(t)$. It is independent of the spectrum of $m(t)$. On the other hand, in PM, $\Delta\omega = k_p m'_p$ depends on the peak value of $\dot{m}(t)$. But $\dot{m}(t)$ depends strongly on the frequency spectrum of $m(t)$. The presence of higher frequency components in $m(t)$ implies rapid time variations, resulting in a higher value of m'_p. Similarly, predominance of lower frequency components will result in a lower value of m'_p. Hence, whereas the WBFM carrier bandwidth [Eq. (5.13)] is practically independent† of the spectrum of $m(t)$, the WBPM carrier bandwidth [Eq. (5.18)] strongly depends on the spectrum of $m(t)$. For $m(t)$ with a spectrum concentrated at lower frequencies, B_{PM} will be smaller than when the spectrum of $m(t)$ is concentrated at higher frequencies.

Verification of FM Bandwidth Relationship

We can verify the bandwidth relations for a specific case of tone modulation; that is, when $m(t)$ is a sinusoid. Let

$$m(t) = \alpha \cos \omega_m t$$

From Eq. (5.6),‡

$$a(t) = \frac{\alpha}{\omega_m} \sin \omega_m t$$

Thus, from Eq. (5.7a), we have

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{\left(j\omega_c t + \frac{k_f \alpha}{\omega_m} \sin \omega_m t\right)}$$

Moreover

$$\Delta\omega = k_f m_p = \alpha k_f$$

and the bandwidth of $m(t)$ is $B = f_m$ Hz. The deviation ratio (or the modulation index, in this case) is

$$\beta = \frac{\Delta f}{f_m} = \frac{\Delta\omega}{\omega_m} = \frac{\alpha k_f}{\omega_m}$$

Hence,

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{j(\omega_c t + \beta \sin \omega_m t)}$$

$$= Ae^{j\omega_c t}(e^{j\beta \sin \omega_m t}) \tag{5.19}$$

---

[*] Equation (5.17a) can be applied only if $m(t)$ is a continuous function of time. If $m(t)$ has jump discontinuities, its derivative does not exist. In such a case, we should use the direct approach (discussed in Example 5.2) to find $\varphi_{\text{PM}}(t)$ and then determine $\Delta\omega$ from $\varphi_{\text{PM}}(t)$.

[†] Except for its weak dependence on B [Eqs. (5.13)].

[‡] Here we are assuming that the constant $a(-\infty) = 0$.

The exponential term in parentheses is a periodic signal with period $2\pi/\omega_m$ and can be expanded by the exponential Fourier series, as usual,

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} C_n e^{j\omega_m t}$$

where

$$C_n = \frac{\omega_m}{2\pi} \int_{-\pi/\omega_m}^{\pi/\omega_m} e^{j\beta \sin \omega_m t} e^{-jn\omega_m t}\, dt$$

Letting $\omega_m t = x$, we get

$$C_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(\beta \sin x - nx)}\, dx$$

The integral on the right-hand side cannot be evaluated in a closed form but must be integrated by expanding the integrand in infinite series. This integral has been extensively tabulated and is denoted by $J_n(\beta)$, the Bessel function of the first kind and nth order. These functions are plotted in Fig. 5.8a as a function of n for various values of β. Thus,

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{jn\omega_m t} \tag{5.20}$$

Substituting Eq. (5.20) into Eq. (5.19), we get

$$\hat{\varphi}_{\text{FM}}(t) = A \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j(\omega_c t + n\omega_m t)}$$

and

$$\hat{\varphi}_{\text{FM}}(t) = A \sum_{n=-\infty}^{\infty} J_n(\beta) \cos{(\omega_c + n\omega_m)t}$$

The modulated signal has a carrier component and an infinite number of sidebands of frequencies $\omega_c \pm \omega_m$, $\omega_c \pm 2\omega_m$, $\ldots$, $\omega_c \pm n\omega_m$, $\ldots$, as shown in Fig. 5.8b. The strength of the nth sideband at $\omega = \omega_c + n\omega_m$ is* $J_n(\beta)$. From the plots of $J_n(\beta)$ in Fig. 5.8a it can be seen that for a given β, $J_n(\beta)$ decreases with n. For a sufficiently large n, $J_n(\beta)$ is negligible, and there are only a finite number of significant sidebands. It can be seen from Fig. 5.8a that $J_n(\beta)$ is negligible for $n > \beta + 1$. Hence, the number of significant sidebands is $\beta + 1$. The bandwidth of the FM carrier is given by

$$\begin{aligned} B_{\text{FM}} &= 2nf_m = 2(\beta + 1)f_m \\ &= 2(\Delta f + B) \end{aligned}$$

which verifies our previous result [Eqs. (5.13)]. When $\beta \ll 1$ (NBFM), there is only one significant sideband and the bandwidth $B_{\text{FM}} = 2f_m = 2B$. It is important to note that this example is a verification, not a proof, of Carson's formula.

---

* Also $J_{-n}(\beta) = (-1)^n J_n(\beta)$. Hence, the magnitude of the LSB at $\omega = \omega_c - n\omega_m$ is the same as that of the USB at $\omega = \omega_c + n\omega_m$.

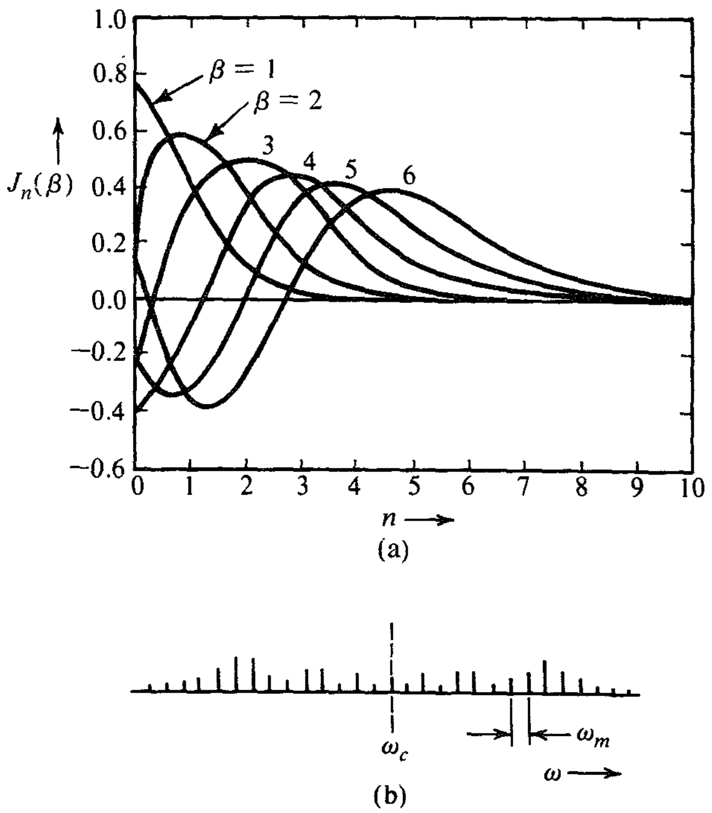


Figure 5.8 (a) Variations of $J_n(\beta)$ as a function of n for various values of β. (b) Tone-modulated FM wave spectrum.

Amplitude modulation is a linear kind of modulation. Hence, most of the results derived for tone modulation are generally valid for other signals. In the literature, tone modulation in FM is often discussed in great details. Unfortunately, angle modulation being a nonlinear kind of modulation, the results derived for tone modulation may have little connection to practical situations. Indeed, these results are meaningless at best and misleading at worst when applied to practical signals. For instance, based on tone modulation analysis, it is often stated that FM is superior to PM by a factor of 3 in terms of the output SNR. We show in Sec. 12.3 that for most of the practical signals, it is PM that is superior to FM. This author feels that too much stress on tone modulation can be misleading. For this reason we have omitted further discussion of tone modulation here.

The method for finding the spectrum of a tone-modulated FM wave can be used for finding the spectrum of an FM wave when $m(t)$ is a general periodic signal. In this case,

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{j\omega_c t}[e^{jk_f a(t)}]$$

Because $a(t)$ is a periodic signal, $e^{jk_f a(t)}$ is also a periodic signal, which can be expressed as an exponential Fourier series in the preceding expression. After this, it is relatively straightforward to write $\hat{\varphi}_{\text{FM}}(t)$ in terms of the carrier and the sidebands.

EXAMPLE 5.3 **(a)** Estimate B_{FM} and B_{PM} for the modulating signal $m(t)$ in Fig. 5.4a for $k_f = 2\pi \times 10^5$ and $k_p = 5\pi$.

(b) Repeat the problem if the amplitude of $m(t)$ is doubled [if $m(t)$ is multiplied by 2].

(a) The peak amplitude of $m(t)$ is unity. Hence, $m_p = 1$. We now determine the essential bandwidth B of $m(t)$. It is left as an exercise for the reader to show that the Fourier series for this periodic signal is given by

$$m(t) = \sum_n C_n \cos n\omega_0 t \qquad \omega_0 = \frac{2\pi}{2 \times 10^{-4}} = 10^4\pi$$

where

$$C_n = \begin{cases} \dfrac{8}{\pi^2 n^2} & n \text{ odd} \\ 0 & n \text{ even} \end{cases}$$

It can be seen that the harmonic amplitudes decrease rapidly with n. The third harmonic is only 11% of the fundamental, and the fifth harmonic is only 4% of the fundamental. This means the third and fifth harmonic powers are 1.21 and 0.16%, respectively, of the fundamental component power. Hence, we are justified in assuming the essential bandwidth of $m(t)$ as the frequency of the third harmonic, that is, $3(10^4/2)$ Hz. Thus,

$$B = 15\,\text{kHz}$$

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = \frac{1}{2\pi}(2\pi \times 10^5)(1) = 100\,\text{kHz}$$

and

$$B_{\text{FM}} = 2(\Delta f + B) = 230\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{100}{15}$$

and

$$B_{\text{FM}} = 2B(\beta + 1) = 30\left(\frac{100}{15} + 1\right) = 230\,\text{kHz}$$

For PM: The peak amplitude of $\dot{m}(t)$, is 20,000, and

$$\Delta f = \frac{1}{2\pi} k_p m'_p = 50\ \text{kHz}$$

Hence,

$$B_{\text{PM}} = 2(\Delta f + B) = 130\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{50}{15}$$

and

$$B_{\text{PM}} = 2B(\beta + 1) = 30\left(\frac{50}{15} + 1\right) = 130\,\text{kHz}$$

(b) Doubling $m(t)$ doubles its peak value. Hence, $m_p = 2$. But its bandwidth is unchanged so that $B = 15$ kHz.

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = \frac{1}{2\pi}(2\pi \times 10^5)(2) = 200\,\text{kHz}$$

and

$$B_{\text{FM}} = 2(\Delta f + B) = 430\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{200}{15}$$

and

$$B_{\text{FM}} = 2B(\beta + 1) = 30\left(\frac{200}{15} + 1\right) = 430\,\text{kHz}$$

For PM: Doubling $m(t)$ doubles its derivative so that now $m_p' = 40,000$, and

$$\Delta f = \frac{1}{2\pi} k_p m_p' = 100\text{kHz}$$

and

$$B_{\text{PM}} = 2(\Delta f + B) = 230\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{100}{15}$$

and

$$B_{\text{PM}} = 2B(\beta + 1) = 30\left(\frac{100}{15} + 1\right) = 230\,\text{kHz}$$

Observe that doubling the signal amplitude [doubling $m(t)$] roughly doubles the bandwidth of both FM and PM waveforms.

EXAMPLE 5.4 Repeat Example 5.3 if $m(t)$ is time-expanded by a factor of 2; that is, if the period of $m(t)$ is 4×10^{-4}.

Recall that time expansion of a signal by a factor of 2 reduces the signal spectral width (bandwidth) by a factor of 2. We can verify this by observing that the fundamental frequency is now 2.5 kHz, and its third harmonic is 7.5 kHz. Hence, $B = 7.5$ kHz, which is half the previous bandwidth. Moreover, time expansion does not affect the peak amplitude so that $m_p = 1$. However, m_p' is halved, that is, $m_p' = 10,000$.

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = 100\,\text{kHz}$$

$$B_{\text{FM}} = 2(\Delta f + B) = 2(100 + 7.5) = 215\,\text{kHz}$$

For PM:

$$\Delta f = \frac{1}{2\pi} k_p m_p' = 25\,\text{kHz}$$

$$B_{\text{PM}} = 2(\Delta f + B) = 65\,\text{kHz}$$

Note that time expansion of $m(t)$ has very little effect on the FM bandwidth, but it halves the PM bandwidth. This verifies our observation that the PM spectrum is strongly dependent on the spectrum of $m(t)$.

EXAMPLE 5.5 An angle-modulated signal with carrier frequency $\omega_c = 2\pi \times 10^5$ is described by the equation

$$\varphi_{\text{EM}}(t) = 10 \cos(\omega_c t + 5 \sin 3000t + 10 \sin 2000\pi t)$$

- **(a)** Find the power of the modulated signal.
- **(b)** Find the frequency deviation Δf.
- **(c)** Find the deviation ratio β.
- **(d)** Find the phase deviation $\Delta\phi$.
- **(e)** Estimate the bandwidth of $\varphi_{\text{EM}}(t)$.

The signal bandwidth is the highest frequency in $m(t)$ (or its derivative). In this case $B = 2000\pi/2\pi = 1000$ Hz.

(a) The carrier amplitude is 10, and the power is

$$P = 10^2/2 = 50$$

(b) To find the frequency deviation Δf, we find the instantaneous frequency ω_i, given by

$$\omega_i = \frac{d}{dt}\theta(t) = \omega_c + 15,000 \cos 3000t + 20,000\pi \cos 2000\pi t$$

The carrier deviation is $15,000 \cos 3000t + 20,000\pi \cos 2000\pi t$. The two sinusoids will add in phase at some point, and the maximum value of this expression is $15,000 + 20,000\pi$. This is the maximum carrier deviation $\Delta\omega$. Hence,

$$\Delta f = \frac{\Delta\omega}{2\pi} = 12,387.32\ \text{Hz}$$

(c)

$$\beta = \frac{\Delta f}{B} = \frac{12,387.32}{1000} = 12.387$$

(d) The angle $\theta(t) = \omega t + (5 \sin 3000t + 10 \sin 2000\pi t)$. The phase deviation is the maximum value of the angle inside the parentheses, and is given by $\Delta\phi = 15$ rad.

(e)

$$B_{EM} = 2(\Delta f + B) = 26,774.65 \text{ Hz}$$

Observe the generality of this method of estimating the bandwidth of an angle-modulated waveform. We need not know whether it is FM, PM, or some other kind of angle modulation. It is applicable to any angle-modulated signal.

A Historical Note: Edwin H. Armstrong (1890–1954)

Today, nobody doubts that FM has a place in broadcasting and communication. As recently as the late sixties, the future of FM broadcasting seemed doomed because of uneconomical operations.

The history of FM is full of strange ironies. The impetus behind the development of FM was the necessity to reduce the transmission bandwidth. Superficial reasoning showed that it was feasible to reduce the transmission bandwidth by using FM. But the experimental results showed otherwise. The transmission bandwidth of FM was actually larger than that of AM. Careful mathematical analysis by Carson showed that FM indeed required a larger bandwidth than AM. Unfortunately, Carson did not recognize the compensating advantage of FM in its ability to suppress noise. Without much basis, he concluded that FM introduced inherent distortion and had no compensating advantages whatsoever.[1] In a later paper he says "In fact, as more and more schemes are analyzed and tested, and as the essential nature of the problem is more clearly perceivable, we are unavoidably forced to the conclusion that static (noise), like the poor, will always be with us."[2] The opinion of one of the ablest mathematicians of the day in the communication field, thus, set back the development of FM by more than a decade. The noise-suppressing advantage of FM was later proved by Major Edwin H. Armstrong,[3] a brilliant engineer whose contributions to the field of radio systems are comparable with those of Hertz and Marconi. It was largely the work of Armstrong that was responsible for rekindling the interest in FM.

Although Armstrong did not invent the concept of FM, he must be considered the father of modern FM. To quote from the early British text *Frequency Modulation Engineering* by Christopher E. Tibbs: "The subject of frequency modulation as we understand it today may be considered to date from Armstrong's paper of 1936. It is true that a good deal of the knowledge of the subject existed prior to that date, but Armstrong was the first to point out in a truly remarkable paper those peculiar characteristics to which modern technique owes its value."[4]

Armstrong was one of the leading architects who laid the groundwork for the mass-communication system. His work on FM came toward the close of his career. Before that, he was well known for several breakthrough contributions to the radio field. *Fortune* magazine says[5]: "Wideband frequency modulation is the fourth, and perhaps the greatest, in a line of Armstrong inventions that have made most of modern broadcasting what it is. Major Armstrong is the acknowledged inventor of the regenerative 'feedback' circuit, which brought radio art out of the crystal-detector headphone stage and made the amplification of broadcasting possible; the superheterodyne circuit, which is the basis of practically all modern radio; and the super-regenerative circuit now in wide use in . . . shortwave systems."

Armstrong was the last of the breed of the lone attic inventors. For the sake of establishing FM broadcasting, he fought a long and costly battle with the radio broadcast establishment, which, abetted by the Federal Communications Commission (FCC), fought tooth and nail to resist FM. In 1944, the FCC, on the basis of erroneous testimony of a technical expert, abruptly shifted the allocated bandwidth of FM from the 42–50-MHz range to 88–108 MHz. This dealt a crippling blow to FM by making obsolete all the equipment (transmitters, receivers, antennas, etc.) that had been built and sold for the old FM bands. Armstrong continued to fight the decision, and in 1947 he succeeded in getting the technical expert to admit his error. In spite of all this, the FCC allocations remained unchanged. Armstrong spent a sizable fortune that he made from previous inventions in legal struggles. The broadcast industry, which so strongly resisted FM, turned around and used his inventions without paying him royalties. Armstrong spent nearly half of his life in the law courts in some of the longest, most notable, and acrimonious patent suits of the era.[4] In the end, with his funds depleted, his energy drained, and his family life shattered, a despondent Armstrong committed suicide (in 1954) by walking out of a window 13 stories above the street.

Features of Angle Modulation

FM (and angle modulation in general) has a number of unique features that recommend it for various radio systems. The transmission bandwidth of AM systems cannot be changed. Because of this AM systems do not have the feature of exchanging signal power for transmission bandwidth. PCM systems have such a feature, and so do angle-modulated systems. In angle modulation, the transmission bandwidth can be adjusted by adjusting Δf. It is shown in Chapter 12 that for angle-modulated systems, the SNR is roughly proportional to the square of the transmission bandwidth B_T. Recall that in PCM, the SNR varies exponentially with B_T and is, therefore, superior to angle modulation.

Immunity of Angle Modulation to Nonlinearities: Another interesting feature of angle modulation is its constant amplitude, which makes it less susceptible to nonlinearities. Consider, for instance, a second-order nonlinear device whose input $x(t)$ and output $y(t)$ are related by

$$y(t) = a_1 x(t) + a_2 x^2(t)$$

If

$$x(t) = \cos\,[\omega_c t + \psi(t)]$$

then

$$\begin{aligned} y(t) &= a_1 \cos\,[\omega_c t + \psi(t)] + a_2 \cos^2[\omega_c t + \psi(t)] \\ &= \frac{a_2}{2} + a_1 \cos\,[\omega_c t + \psi(t)] + \frac{a_2}{2} \cos\,[2\omega_c t + 2\psi(t)] \end{aligned}$$

For the FM wave $\psi(t) = k_f \int m(\alpha)\,d\alpha$, and

$$y(t) = \frac{a_2}{2} + a_1 \cos\left[\omega_c t + k_f \int m(\alpha)\,d\alpha\right] + \frac{a_2}{2} \cos\left[2\omega_c t + 2k_f \int m(\alpha)\,d\alpha\right]$$

The dc term is filtered out to give the output that contains the original signal plus an additional FM signal, whose carrier frequency as well as frequency deviation are multiplied by 2. Note, however, that the information $m(t)$ is intact in both terms. Thus, the nonlinearity has not distorted the information in any way. Because of the property of multiplying the carrier frequency, such nonlinear devices are also called **frequency multipliers**.

In the preceding case, because the device was of second order, it multiplied the frequency by 2. We can generalize this result for an nth-order multiplier (nonlinear device). Any nonlinear device, such as a diode or a transistor, can be used for this purpose. The characteristic of these devices can be expressed as

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + \cdots + a_n x^n(t) \tag{5.21}$$

If $x(t) = A \cos\left[\omega_c t + k_f \int m(\alpha)\,d\alpha\right]$, then using trigonometric identities, we can readily show that $y(t)$ is of the form

$$\begin{aligned} y(t) = c_o + c_1 \cos\left[\omega_c t + k_f \int m(\alpha)\,d\alpha\right] + c_2 \cos\left[2\omega_c t + 2k_f \int m(\alpha)\,d\alpha\right] \\ + \cdots + c_n \cos\left[n\omega_c t + nk_f \int m(\alpha)\,d\alpha\right] \end{aligned} \tag{5.22}$$

Hence, the output will have spectra at ω_c, $2\omega_c, \ldots,$ $n\omega_c$, with frequency deviations Δf, $2\Delta f$, $\ldots$, $n\Delta f$, respectively. Hence, the nonlinearity generates components at unwanted frequencies. But the desired term $\cos\,[\omega_c t + \psi(t)]$ is undistorted, and by using a bandpass filter centered at ω_c, we can suppress all unwanted terms in $y(t)$ and obtain the desired signal component without distortion. Note that even the unwanted terms have the desired information intact, and any one of the unwanted terms can be used to extract information. The term $\cos\,[2\omega_c t + 2k_f \int m(\alpha)\,d\alpha]$, for instance, has twice the original carrier frequency and twice the original frequency deviation. Hence, such nonlinear devices can be used to increase the carrier frequency as well as the frequency deviation.

A similar nonlinearity in AM not only causes unwanted modulation with carrier frequencies $n\omega_c$ but also causes distortion of the desired signal. For instance, if a DSB-SC signal $m(t) \cos \omega_c t$ passes through a nonlinearity $y(t) = a\,x(t) + b\,x^3(t)$, the output is

$$\begin{aligned} y(t) &= am(t) \cos \omega_c t + bm^3(t) \cos^3 \omega_c t \\ &= \left[am(t) + \frac{3b}{4} m^3(t)\right] \cos \omega_c t + \frac{b}{4} m^3(t) \cos 3\omega_c t \end{aligned}$$

Passing this signal through a bandpass filter yields $[am(t) + (3b/4)m^3(t)] \cos \omega_c t$. Observe the distortion component $(3b/4)m^3(t)$ present along with the desired signal $am(t)$.

Immunity from nonlinearity is the primary reason why angle modulation is used in microwave radio relay systems, where power levels are high. This requires highly efficient nonlinear class C amplifiers. In addition, the constant amplitude of FM gives it a kind of immunity against rapid fading. The effect of amplitude variations caused by rapid fading can be eliminated by using automatic gain control and bandpass limiting (discussed in Sec. 5.4). These features make FM attractive for microwave radio relay systems. Angle modulation is also less vulnerable than AM to small signal interference from adjacent channels. Finally, as stated earlier, FM is capable of exchanging SNR for the transmission bandwidth.

In telephone systems, several channels are multiplexed using SSB signals. The multiplexed signal is frequency modulated and transmitted over a microwave radio relay system with many links in tandem. In this application, however, FM is used not to realize the noise reduction but to realize other advantages of constant amplitude, and, hence, NBFM rather than WBFM is used.

WBFM is used widely in space and satellite communication systems. The large bandwidth expansion reduces the required SNR and thus reduces the transmitter power requirement —which is very important because of weight considerations in space. WBFM is also used for high-fidelity radio transmission over rather limited areas.

5.3 GENERATION OF FM WAVES

Basically, there are two ways of generating FM waves: **indirect generation** and **direct generation**.

Indirect Method of Armstrong

In this method, NBFM is generated by integrating $m(t)$ and using it to phase modulate a carrier, as shown in Fig. 5.6b [or Eq. (5.9)]. The NBFM is then converted to WBFM by using frequency multipliers (discussed earlier), as shown in Fig. 5.9. Thus, if we want a 12-fold increase in the frequency deviation, we can use a 12th-order nonlinear device or two second-order and one third-order devices in cascade. The output has a bandpass filter centered at $12\omega_c$, so that it selects only the appropriate term, whose carrier frequency as well as the frequency deviation Δf are 12 times the original values. Generally, we require to increase Δf by a very large factor n. This increases the carrier frequency also by n. Such a large increase in the carrier frequency may not be needed. In this case we can use frequency mixing (see Example 4.2, Fig. 4.7) to shift down the carrier frequency to the desired value (recall that a frequency mixer shifts the carrier frequency).

The NBFM generated by Armstrong's method (Fig. 5.6b) has some distortion because of the approximation of Eqs. (5.8) by Eq. (5.9) (see Example 5.6). The output of the Armstrong NBFM modulator, as a result, also has some amplitude modulation. Amplitude limiting in the frequency multipliers removes most of this distortion.

A simplified diagram of a commercial FM transmitter using Armstrong's method is shown in Fig. 5.10. The final output is required to have a carrier frequency of 91.2 MHz and $\Delta f = 75$ kHz. We begin with NBFM with a carrier frequency $f_{c_1} = 200$ kHz generated by a crystal oscillator. This frequency is chosen because it is easy to construct stable crystal oscillators as well as balanced modulators at this frequency. The deviation Δf is chosen to be 25 Hz in order to maintain $\beta \ll 1$, as required in NBPM. For tone modulation $\beta = \Delta f / f_m$. The baseband spectrum (required for high-fidelity purposes) ranges from 50 Hz to 15 kHz. The choice of $\Delta f = 25$ Hz is reasonable because it gives $\beta = 0.5$ for the worst possible case ($f_m = 50$).

In order to achieve $\Delta f = 75$ kHz, we need a multiplication of 75,000/25 = 3000. This can be done by two multiplier stages, of 64 and 48, as shown in Fig. 5.10, giving a

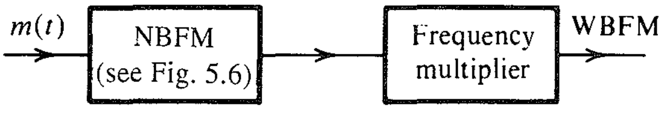


Figure 5.9 Simplified block diagram of Armstrong indirect FM wave generator.

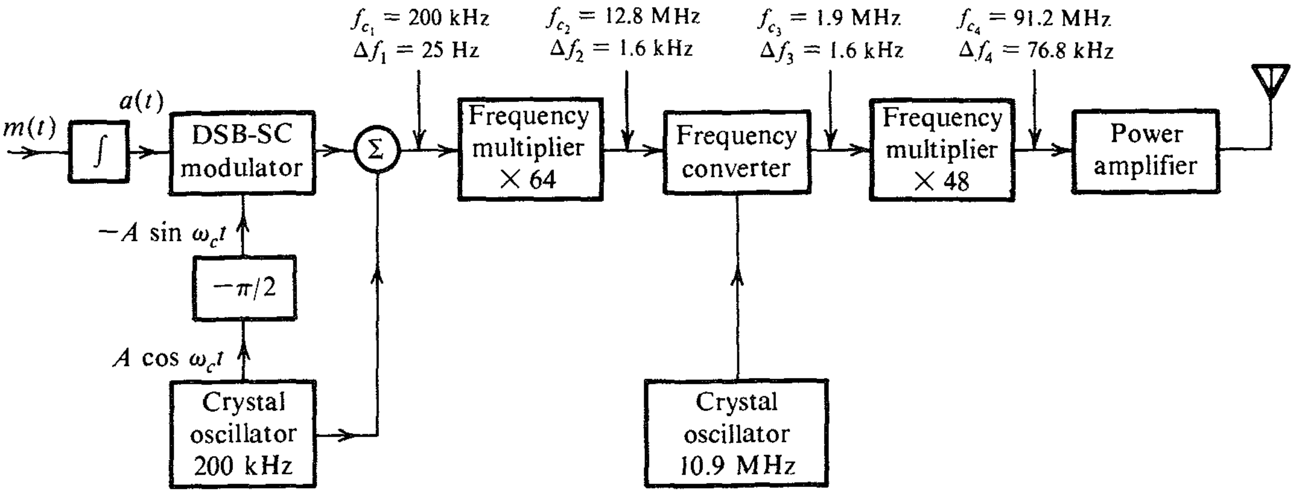

$f_{c_1} = 200$ kHz
$\Delta f_1 = 25$ Hz
$f_{c_2} = 12.8$ MHz
$\Delta f_2 = 1.6$ kHz
$f_{c_3} = 1.9$ MHz
$\Delta f_3 = 1.6$ kHz
$f_{c_4} = 91.2$ MHz
$\Delta f_4 = 76.8$ kHz
$m(t)$
$a(t)$
$\int$
DSB-SC modulator
Σ
Frequency multiplier × 64
Frequency converter
Frequency multiplier × 48
Power amplifier
$-A \sin \omega_c t$
$-\pi/2$
$A \cos \omega_c t$
Crystal oscillator 200 kHz
Crystal oscillator 10.9 MHz

Figure 5.10 Armstrong indirect FM transmitter.

total multiplication of $64 \times 48 = 3072$, and $\Delta f = 76.8$ kHz.* The multiplication is effected by using frequency doublers and triplers in cascade, as needed. Thus, a multiplication of 64 can be obtained by six doublers in cascade, and a multiplication of 48 can be obtained by four doublers and a tripler in cascade. Multiplication of $f_c = 200$ kHz by 3072, however, would yield a final carrier of about 600 MHz. This difficulty is avoided by using a frequency translation, or conversion, after the first multiplier (Fig. 5.10). The first multiplication by 64 results in the carrier frequency $f_{c_2} = 200\,\text{kHz} \times 64 = 12.8$ MHz, and the carrier deviation $\Delta f_2 = 25 \times 64 = 1.6$ kHz. We now shift the entire spectrum using a frequency converter (or mixer) with carrier frequency 10.9 MHz. This results in a new carrier frequency $f_{c_3} = 12.8 - 10.9 = 1.9$ MHz. The frequency converter shifts the entire spectrum without altering Δf. Hence, $\Delta f_3 = 1.6$ kHz. Further multiplication, by 48, yields $f_{c_4} = 1.9 \times 48 = 91.2$ MHz and $\Delta f_4 = 1.6 \times 48 = 76.8$ kHz.

This scheme has an advantage of frequency stability, but it suffers from inherent noise caused by excessive multiplication and distortion at lower modulating frequencies, where $\Delta f / f_m$ is not small enough.

EXAMPLE 5.6 Discuss the nature of distortion inherent in the Armstrong indirect FM generator.

Two kinds of distortions arise in this scheme: amplitude distortion and frequency distortion. The NBFM wave is given by [Eq. (5.9)]

$$\begin{aligned}\varphi_{\text{FM}}(t) &= A[\cos\omega_c t - k_f a(t)\sin\omega_c t] \\ &= AE(t)\cos[\omega_c t + \theta(t)]\end{aligned}$$

where

$$E(t) = \sqrt{1 + k_f^2 a^2(t)} \quad \text{and} \quad \theta(t) = \tan^{-1}[k_f a(t)]$$

---

* If we wish Δf to be exactly 75 kHz instead of 76.8 kHz, we must reduce the narrow-band Δf from 25 Hz to $25(75/76.8) = 24.41$ Hz.

Amplitude distortion occurs because the amplitude $AE(t)$ of the modulated waveform is not constant. This is not a serious problem, because amplitude variations can be eliminated by a bandpass limiter discussed in the next section (see Fig. 5.12). Ideally, $\theta(t)$ should be $k_f a(t)$. Instead, the phase $\theta(t)$ in the preceding equation is

$$\theta(t) = \tan^{-1}[k_f a(t)]$$

and the instantaneous frequency $\omega_i(t)$ is

$$\begin{aligned}\omega_i(t) = \dot{\theta}(t) &= \frac{k_f \dot{a}(t)}{1 + k_f^2 a^2(t)} \\ &= \frac{k_f m(t)}{1 + k_f^2 a^2(t)} \\ &= k_f m(t)[1 - k_f^2 a^2(t) + k_f^4 a^4(t) - \cdots]\end{aligned}$$

Ideally, the instantaneous frequency should be $k_f m(t)$. The remaining terms in this equation are the distortion.

Let us investigate the effect of this distortion in tone modulation, where $m(t) = \alpha \cos \omega_m t$, $a(t) = \alpha \sin \omega_m t / \omega_m$, and the modulation index $\beta = \alpha k_f / \omega_m$. Hence,

$$\omega_i(t) = \beta \omega_m \cos \omega_m t (1 - \beta^2 \sin^2 \omega_m t + \beta^4 \sin^4 \omega_m t - \cdots)$$

It is evident from this equation that this scheme has odd-harmonic distortion, the most important term being the third harmonic. Ignoring the remaining terms, this equation becomes

$$\begin{aligned}\omega_i(t) &\simeq \beta \omega_m \cos \omega_m t (1 - \beta^2 \sin^2 \omega_m t) \\ &= \beta \omega_m \left(1 - \frac{\beta^2}{4}\right) \cos \omega_m t + \frac{\beta^3 \omega_m}{4} \cos 3\omega_m t \\ &\simeq \underbrace{\beta \omega_m \cos \omega_m t}_{\text{desired}} + \underbrace{\frac{\beta^3 \omega_m}{4} \cos 3\omega_m t}_{\text{distortion}} \qquad \text{for } \beta \ll 1\end{aligned}$$

The ratio of the third harmonic distortion to the desired signal is $\beta^2/4$. For the generator in Fig. 5.10, the worst possible case occurs at the lower modulation frequency of 50 Hz, where $\beta = 0.5$. In this case the third harmonic distortion is 1/16, or 6.25%.

Direct Generation

In a voltage-controlled oscillator (VCO), the frequency is controlled by an external voltage. The oscillation frequency varies linearly with the control voltage. We can generate an FM wave by using the modulating signal $m(t)$ as a control signal. This gives

$$\omega_i(t) = \omega_c + k_f m(t)$$

One can construct a VCO using an operational amplifier and an hysteric comparator[6] (such as a Schmitt trigger circuit). Another way of accomplishing the same goal is to vary one of the reactive parameters (C or L) of the resonant circuit of an oscillator. A reverse-biased semiconductor diode acts as a capacitor whose capacitance varies with the bias voltage. The capacitance of these diodes, known under several trade names (such as varicaps, varactors, or voltacaps), can be approximated as a linear function of the bias voltage $m(t)$ over a limited range. In Hartley or Colpitt oscillators, for instance, the frequency of oscillation is given by

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

If the capacitance C is varied by the modulating signal $m(t)$, that is, if

$$\begin{aligned} C &= C_0 - km(t) \\ \omega_0 &= \frac{1}{\sqrt{LC_0\left[1 - \dfrac{km(t)}{C_0}\right]}} \\ &= \frac{1}{\sqrt{LC_0}\left[1 - \dfrac{km(t)}{C_0}\right]^{1/2}} \\ &\approx \frac{1}{\sqrt{LC_0}}\left[1 + \frac{km(t)}{2C_0}\right] \qquad \frac{km(t)}{C_0} \ll 1 \end{aligned}$$

Here we have used the binomial approximation $(1+x)^n \approx 1 + nx$ for $|x| \ll 1$. Thus,

$$\begin{aligned} \omega_0 &= \omega_c\left[1 + \frac{km(t)}{2C_0}\right] \qquad & \omega_c &= \frac{1}{\sqrt{LC_0}} \\ &= \omega_c + k_f m(t) \qquad & k_f &= \frac{k\omega_c}{2C_0} \end{aligned}$$

Because $C = C_0 - km(t)$, the maximum capacitance deviation is

$$\Delta C = km_p = \frac{2k_f C_0 m_p}{\omega_c}$$

Hence,

$$\frac{\Delta C}{C_0} = \frac{2k_f m_p}{\omega_c} = \frac{2\Delta f}{f_c}$$

In practice, $\Delta f / f_c$ is usually small, and, hence, ΔC is a small fraction of C_0, which helps limit the harmonic distortion that arises because of the approximation used in this derivation.

We may also generate direct FM by using a saturable core reactor, where the inductance of a coil is varied by a current through a second coil (also wound around the same core). This results in a variable inductor whose inductance is proportional to the current in the second coil.

Direct FM generation generally produces sufficient frequency deviation and requires little frequency multiplication. But this method has poor frequency stability. In practice, feedback is used to stabilize the frequency. The output frequency is compared with a constant frequency generated by a stable crystal oscillator. An error signal (error in frequency) is detected and fed back to the oscillator to correct the error.

5.4 DEMODULATION OF FM

The information in an FM signal resides in the instantaneous frequency $\omega_i = \omega_c + k_f m(t)$. Hence, a frequency-selective network with a transfer function of the form $|H(\omega)| = a\omega + b$ over the FM band would yield an output proportional to the instantaneous frequency (Fig. 5.11a).* There are several possible networks with such characteristics. The simplest among them is an ideal differentiator with the transfer function $j\omega$.

If we apply $\varphi_{\text{FM}}(t)$ to an ideal differentiator, the output is

$$\begin{aligned}\dot{\varphi}_{\text{FM}}(t) &= \frac{d}{dt}\left\{A \cos\left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha\right]\right\} \\ &= A\left[\omega_c + k_f m(t)\right] \sin\left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d(\alpha)\right] \qquad (5.23)\end{aligned}$$

The signal $\dot{\varphi}_{\text{FM}}(t)$ is both amplitude and frequency modulated (Fig. 5.11b), the envelope being $A[\omega_c + k_f m(t)]$. Because $\Delta\omega = k_f m_p < \omega_c$, $\omega_c + k_f m(t) > 0$ for all t, and $m(t)$ can be obtained by envelope detection of $\dot{\varphi}_{\text{FM}}(t)$ (Fig. 5.11c).

The amplitude A of the incoming FM carrier is assumed to be constant. If the amplitude A were not constant, but a function of time, there would be an additional term containing

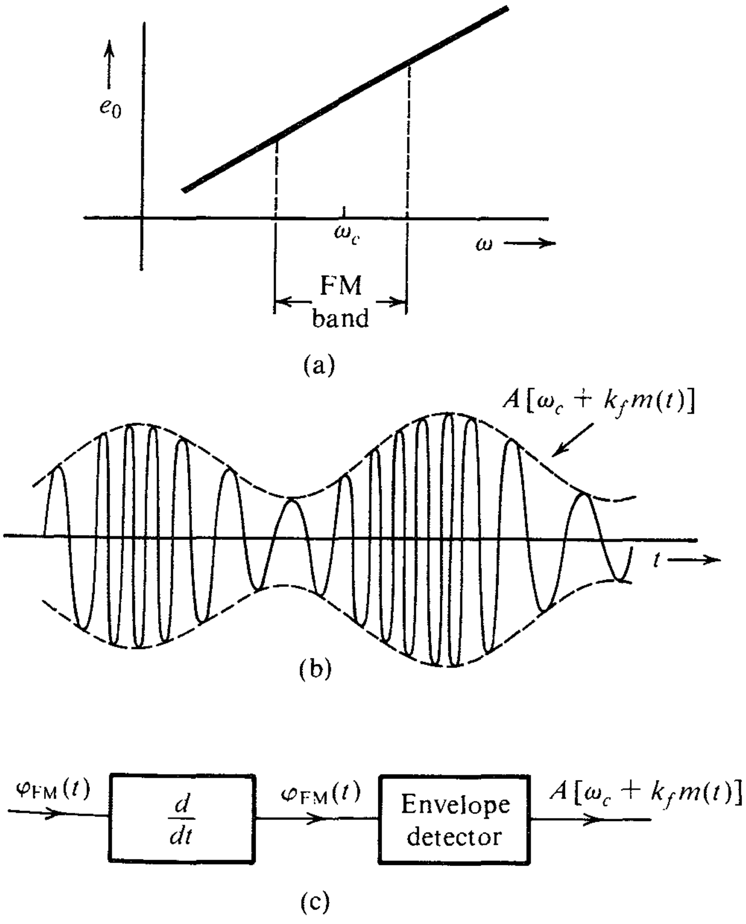


Figure 5.11 (a) FM demodulator frequency response. (b) Output of a differentiator to the input FM wave. (c) FM demodulation by direct differentiation.

---

* Provided the variations of ω_i are slow in comparison to the time constant of the network.

dA/dt on the right-hand side of Eq. (5.23). Even if this term were neglected, the envelope of $\varphi_{\text{FM}}(t)$ would be $A(t)[\omega_c + k_f m(t)]$, and the envelope-detector output would be proportional to $m(t)A(t)$. Hence, it is essential to maintain A constant. Several factors, such as channel noise, fading, and so on, cause A to vary. This variation in A should be removed before applying the signal to the FM detector.

Bandpass Limiter

The amplitude variations of an angle-modulated carrier can be eliminated by what is known as a **bandpass limiter**, which consists of a hard limiter followed by a bandpass filter (Fig. 5.12a). The input-output characteristic of a hard limiter is shown in Fig. 5.12b. Observe that the bandpass limiter output to a sinusoid will be a square wave of unit amplitude regardless of the incoming sinusoidal amplitude. Moreover, the zero crossings of the incoming sinusoid are preserved in the output because when the input is zero, the output is also zero (Fig. 5.12b). Thus an angle-modulated sinusoidal input $v_i(t) = A(t) \cos \theta(t)$ results in a constant-amplitude, angle-modulated square wave $v_o(t)$, as shown in Fig. 5.12c. As we have seen earlier, such a nonlinear operation preserves the angle modulation information. When $v_o(t)$ is passed through a bandpass filter centered at ω_c, the output is a constant-amplitude, angle-modulated wave. To show this, consider the incoming angle-modulated wave

$$v_i(t) = A(t) \cos \theta(t)$$

where

$$\theta(t) = \omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha$$

The output $v_o(t)$ of the hard limiter is $+1$ or -1, depending on whether $v_i(t) = A(t) \cos \theta(t)$ is positive or negative (Fig. 5.12c). Because $A(t) \geq 0$, $v_o(t)$ can be expressed as a function of θ:

$$v_o(\theta) = \begin{cases} 1 & \cos\theta > 0 \\ -1 & \cos\theta < 0 \end{cases}$$

Hence, v_o as a function of θ is a periodic square-wave function with period 2π (Fig. 5.12d), which can be expanded by a Fourier series [see Eq. (2.76)],

$$v_o(\theta) = \frac{4}{\pi}\left(\cos\theta - \frac{1}{3}\cos 3\theta + \frac{1}{5}\cos 5\theta + \cdots\right)$$

This is valid for any real variable θ. At any instant t, $\theta = \omega_c t + k_f \int m(\alpha)\, d\alpha$, and the output is $v_o[\omega_c t + k_f \int m(\alpha)\, d\alpha]$. Hence, the output v_o as a function of time is given by

$$\begin{aligned} v_o[\theta(t)] &= v_o\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \\ &= \frac{4}{\pi}\left\{\cos\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] - \frac{1}{3}\cos 3\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \right. \\ &\quad \left. + \frac{1}{5}\cos 5\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \cdots\right\} \end{aligned}$$

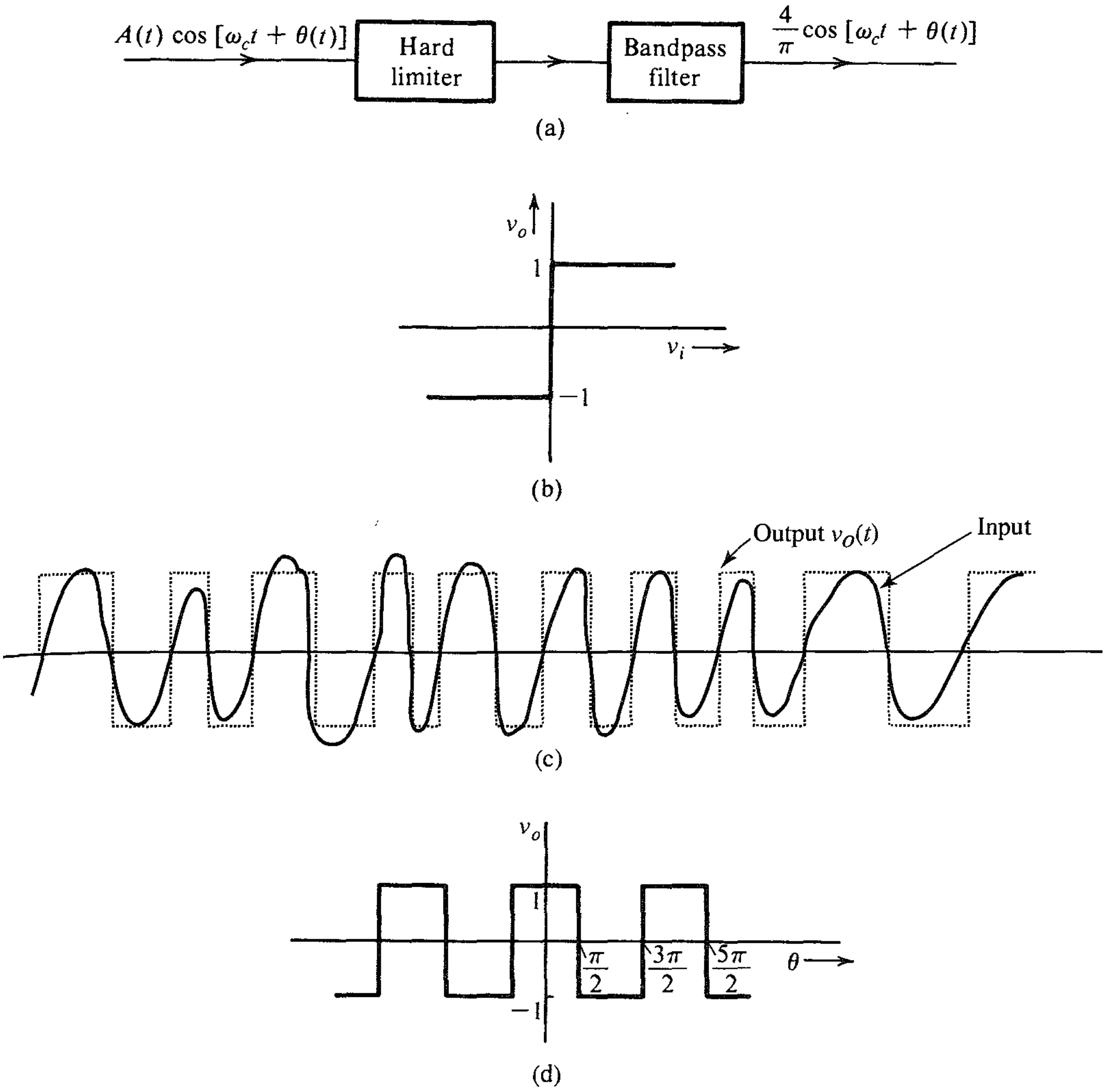

$A(t)\cos[\omega_c t + \theta(t)]$
Hard limiter
Bandpass filter
$\frac{4}{\pi}\cos[\omega_c t + \theta(t)]$
(a)
v_o
1
v_i
−1
(b)
Output $v_o(t)$
Input
(c)
v_o
1
$\frac{\pi}{2}$
$\frac{3\pi}{2}$
$\frac{5\pi}{2}$
θ
−1
(d)

Figure 5.12 (a) Hard limiter and bandpass filter used to remove amplitude variations in FM wave. (b) Hard limiter input-output characteristic. (c) Hard limiter input and the corresponding output. (d) Hard limiter output as a function of θ.

The output, therefore, has the original FM wave plus a frequency-multiplied FM wave with multiplication factors of 3, 5, 7, ... We can pass the output of the hard limiter through a bandpass filter with a center frequency ω_c and a bandwidth B_{FM}, as shown in Fig. 5.12a. The filter output $e_o(t)$ is the desired angle-modulated carrier with a constant amplitude,

$$e_o(t) = \frac{4}{\pi}\cos\left[\omega_c(t) + k_f \int m(\alpha)d\alpha\right]$$

Although we derived these results for FM, this applies to PM (angle modulation in general) as well. The bandpass filter not only maintains the constant amplitude of the angle-modulated carrier but also partially suppresses the channel noise when the noise is small.[7]

Practical Frequency Demodulators

One can use an operational amplifier differentiator as an FM demodulator. A simple tuned circuit followed by an envelope detector can also serve as a frequency detector because its frequency response $|H(\omega)|$ below (or above) the resonance frequency is approximately linear of the form $a\omega + b$. Since the operation is on the slope of $|H(\omega)|$, this method is also called **slope detection**. It suffers from the fact that the slope of $|H(\omega)|$ is linear over only a small band and, hence, causes considerable distortion in the output. This fault can be partially corrected by a **balanced discriminator**.

Another balanced demodulator, the **ratio detector**, also widely used in the past, offers better protection against carrier amplitude variations than does the discriminator. For many years ratio detectors were standard in almost all FM receivers.^{8}

Zero-crossing detectors are also used because of advances in digital integrated circuits. These are the **frequency counters** designed to measure the instantaneous frequency by the number of zero crossings. The rate of zero crossings is equal to the instantaneous frequency of the input signal.

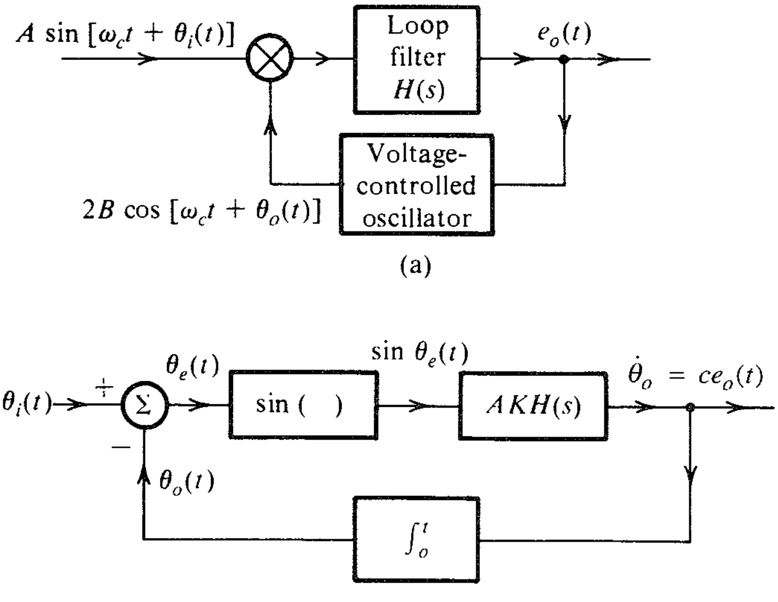
Phase-Locked Loop (PLL): Because of their low cost and superior performance, especially when the SNR is low, FM demodulation using PLL is the most widely used method today. In Chapter 4, we saw how a PLL tracks the incoming signal angle and instantaneous frequency. Consider the PLL in Fig. 5.13a. The output $e_o(t)$ of the loop filter $H(s)$ acts as an input to the VCO (Fig. 5.13a). The free-running frequency of VCO is set at the carrier frequency ω_c. The instantaneous frequency of the VCO is given by [see Eq. (4.25)]

$$\omega_{\text{VCO}} = \omega_c + ce_o(t)$$

If the VCO output is $B\cos[\omega_c t + \theta_o(t)]$, then its instantaneous frequency is $\omega_c + \dot{\theta}_o(t)$. Therefore,

$$\dot{\theta}_o(t) = ce_o(t) \tag{5.24}$$

where c and B are constants of the PLL.


$A\sin[\omega_c t + \theta_i(t)]$
Loop filter $H(s)$
$e_o(t)$
Voltage-controlled oscillator
$2B\cos[\omega_c t + \theta_o(t)]$
$\theta_i(t)$
+
−
Σ
$\theta_e(t)$
sin ( )
$\sin\theta_e(t)$
$AKH(s)$
$\dot{\theta}_o = ce_o(t)$
$\theta_o(t)$
$\int_o^t$

(a)

(b)

Figure 5.13 Phase-locked loop and its equivalent circuit.

Let the incoming signal (input to the PLL) be $A \sin[\omega_c t + \theta_i(t)]$. If the incoming signal happens to be $A \sin[\omega_o t + \psi(t)]$, it can still be expressed as $A \sin[\omega_c t + \theta_i(t)]$, where $\theta_i(t) = (\omega_o - \omega_c)t + \psi(t)$. Hence, the analysis that follows is general and not restricted to equal frequencies of the incoming signal and the free-running VCO signal.

The multiplier output is

$$AB \sin(\omega_c t + \theta_i)\cos(\omega_c t + \theta_o) = \frac{AB}{2}[\sin(\theta_i - \theta_o) + \sin(2\omega_c t + \theta_i + \theta_o)]$$

The sum frequency term is suppressed by the loop filter. Hence, the effective input to the loop filter is $\frac{1}{2}AB \sin[\theta_i(t) - \theta_o(t)]$. If $h(t)$ is the unit impulse response of the loop filter,

$$\begin{aligned} e_o(t) &= h(t) * \frac{1}{2}AB \sin[\theta_i(t) - \theta_o(t)] \\ &= \frac{1}{2}AB \int_0^t h(t-x) \sin[\theta_i(x) - \theta_o(x)]\, dx \qquad (5.25) \end{aligned}$$

Substituting Eq. (5.24) in Eq. (5.25),

$$\dot{\theta}_o(t) = AK \int_0^t h(t-x) \sin \theta_e(x)\, dx \qquad (5.26)$$

where $K = \frac{1}{2}cB$ and $\theta_e(t)$ is the phase error, defined as

$$\theta_e(t) = \theta_i(t) - \theta_o(t)$$

These equations [along with Eq. (5.24)] immediately suggest a model for the PLL, as shown in Fig. 5.13b.

When the incoming FM carrier* is $A \sin[\omega_c t + \theta_i(t)]$,

$$\theta_i(t) = k_f \int_{-\infty}^t m(\alpha)\, d\alpha \qquad (5.27)$$

Hence,

$$\theta_o(t) = k_f \int_{-\infty}^t m(\alpha)\, d\alpha - \theta_e$$

and, assuming a small error θ_e,

$$e_o(t) = \frac{1}{c}\dot{\theta}_o(t) \simeq \frac{k_f}{c} m(t) \qquad (5.28)$$

Thus, the PLL acts as an FM demodulator. If the incoming signal is a PM wave, $\theta_o(t) = \theta_i(t) = k_p m(t)$ and $e_o(t) = k_p \dot{m}(t)/c$. In this case we need to integrate $e_o(t)$ to obtain the desired signal. A detailed analysis of PLL is given next for two special cases.

Small-Error Analysis

In this case, $\sin\theta_e \simeq \theta_e$, and the block diagram in Fig. 5.13b reduces to the linear (time-invariant) system shown in Fig. 5.14a. Straightforward calculation gives

---

* Here we are using $\sin[\omega_c t + \theta_i(t)]$ rather than the usual $\cos[\omega_c t + \theta_i(t)]$. This is really immaterial, because a cosine can be expressed as a sine with a $\pi/2$ phase addition. Because the final step [Eq. (5.28)] involves differentiation of the angle, the constant phase vanishes.

$$\frac{\Theta_o(s)}{\Theta_i(s)} = \frac{AKH(s)/s}{1 + [AKH(s)/s]} = \frac{AKH(s)}{s + AKH(s)} \tag{5.29}$$

Therefore, the PLL acts as a filter with transfer function $AKH(s)/[s + AKH(s)]$, as shown in Fig. 5.14b. The error $\Theta_e(s)$ is given by

$$\begin{aligned}\Theta_e(s) &= \Theta_i(s) - \Theta_o(s) = \left[1 - \frac{\Theta_o(s)}{\Theta_i(s)}\right]\Theta_i(s) \\ &= \frac{s}{s + AKH(s)}\Theta_i(s)\end{aligned} \tag{5.30}$$

One of the important applications of the PLL is in the acquisition of the frequency and the phase for the purpose of synchronization. Let the incoming signal be $A\ \sin(\omega_0 t + \varphi_0)$. We wish to generate a local signal of frequency ω_0 and phase* φ_0. Assuming the quiescent frequency of the VCO to be ω_c, the incoming signal can be expressed as $A\ \sin[\omega_c t + \theta_i(t)]$, where

$$\theta_i(t) = (\omega_0 - \omega_c)t + \varphi_0$$

and

$$\Theta_i(s) = \frac{(\omega_0 - \omega_c)}{s^2} + \frac{\varphi_0}{s}$$

Consider the special case of $H(s) = 1$. Substituting this equation into Eq. (5.30),

$$\begin{aligned}\Theta_e(s) &= \frac{s}{s + AK}\left[\frac{\omega_0 - \omega_c}{s^2} + \frac{\varphi_0}{s}\right] \\ &= \frac{(\omega_o - \omega_c)/AK}{s} - \frac{(\omega_0 - \omega_c)/AK}{s + AK} + \frac{\varphi_0}{s + AK}\end{aligned}$$

Hence,

$$\theta_e(t) = \frac{(\omega_0 - \omega_c)}{AK}\left(1 - e^{-AKt}\right) + \varphi_0 e^{-AKt} \tag{5.31a}$$

Observe that

$$\lim_{t\to\infty} \theta_e(t) = \frac{\omega_0 - \omega_c}{AK} \tag{5.31b}$$

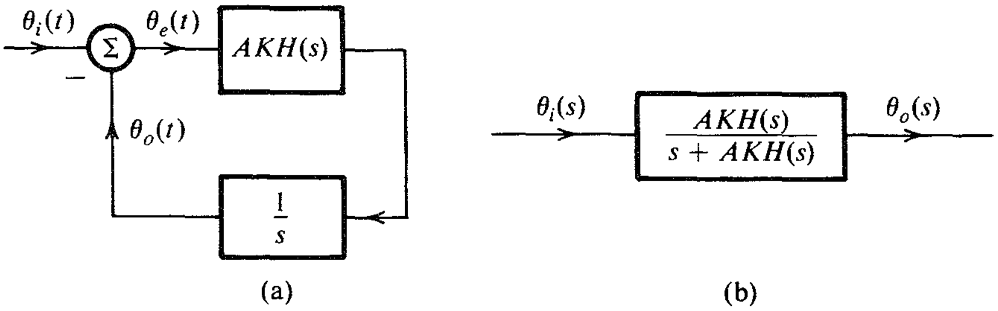


Figure 5.14 Equivalent circuits of a linearized PLL.

---

* With a difference $\pi/2$.

Hence, after the transient dies (in about $4/AK$ seconds), the phase error maintains a constant value of $(\omega_0 - \omega_c)/AK$. This means the PLL frequency eventually equals the incoming frequency ω_0. There is, however, a constant phase error. The PLL output is

$$B \cos \left[\omega_0 t + \varphi_0 - \frac{\omega_0 - \omega_c}{AK}\right]$$

For a second-order PLL using

$$H(s) = \frac{s+a}{s} \tag{5.32a}$$

$$\Theta_e(s) = \frac{s}{s + AKH(s)}\Theta_i(s) \tag{5.32b}$$

$$= \frac{s^2}{s^2 + AK(s+a)}\left[\frac{\omega_0 - \omega_c}{s^2} + \frac{\varphi_0}{s}\right]$$

The final-value theorem directly yields[9]

$$\lim_{t\to\infty} \theta_e(t) = \lim_{s\to 0} s\Theta_e(s) = 0 \tag{5.33}$$

In this case, the PLL eventually acquires both the frequency and the phase of the incoming signal.

Using small-error analysis, it can be shown that a first-order loop cannot track an incoming signal whose instantaneous frequency is varying linearly with time. Moreover, such a signal can be tracked within a constant phase (constant phase error) by using a second-order loop [Eq. (5.32)], and it can be tracked with zero phase error using a third-order loop.[10]

It must be remembered that the preceding analysis assumes a linear model, which is valid only when $\theta_e(t) \ll \pi/2$. This means the frequencies ω_0 and ω_c must be very close for this analysis to be valid. For a general case, one must use the nonlinear model in Fig. 5.13b. For such an analysis, the reader is referred to Viterbi[10] or Lindsey.[11]

To analyze PLL behavior as an FM demodulator, we consider the case of a small error (linear model of the PLL) with $H(s) = 1$. For this case, Eq. (5.29) becomes

$$\Theta_o(s) = \frac{AK}{s + AK}\Theta_i(s)$$

If $E_o(s)$ and $M(s)$ are Fourier transforms of $e_o(t)$ and $m(t)$, respectively, then from Eqs. (5.27) and (5.28) we have

$$\Theta_i(s) = \frac{k_f M(s)}{s} \qquad \text{and} \qquad s\Theta_o(s) = cE_o(s)$$

Hence,

$$E_o(s) = \left(\frac{k_f}{c}\right)\frac{AK}{s + AK}M(s)$$

Thus, the PLL output $e_o(t)$ is a distorted version of $m(t)$ and is equivalent to the output of a single-pole circuit (such as a simple RC circuit) with transfer function $k_f AK/c(s + AK)$ with $m(t)$ as the input. To reduce distortion, we must choose AK well above the radian bandwidth of $m(t)$, so that $e_o(t) \simeq k_f m(t)/c$.

In the presence of small noise, the behavior of the PLL is comparable to that of a frequency discriminator. The advantage of the PLL over a frequency discriminator appears only when the noise is large.

First-Order-Loop Analysis

Here we shall use the nonlinear model in Fig. 5.13b, but for the simple case of $H(s) = 1$. For this case $h(t) = \delta(t)$,* and Eq. (5.26) gives

$$\dot{\theta}_o(t) = AK \sin \theta_e(t)$$

Because $\theta_e = \theta_i - \theta_o$,

$$\dot{\theta}_e = \dot{\theta}_i - AK \sin \theta_e(t) \tag{5.34}$$

Let us here consider the problem of frequency and phase acquisition. Let the incoming signal be $A \sin (\omega_0 t + \varphi_0)$, and the VCO has a quiescent frequency ω_c. Hence,

$$\theta_i(t) = (\omega_0 - \omega_c)t + \varphi_0$$

and

$$\dot{\theta}_e = (\omega_0 - \omega_c) - AK \sin \theta_e(t) \tag{5.35}$$

For a better understanding of the PLL behavior, we use Eq. (5.35) to sketch $\dot{\theta}_e$ vs. θ_e. Equation (5.35) shows that $\dot{\theta}_e$ is a vertically shifted sinusoid, as shown in Fig. 5.15. To satisfy Eq. (5.35), the loop operation must stay along the sinusoidal trajectory shown in Fig. 5.15. When $\dot{\theta}_e = 0$, the system is in equilibrium, because at these points, θ_e stops varying with time. Thus $\theta_e = \theta_1$, θ_2, θ_3, and θ_4 are all equilibrium points.

If the initial phase error $\theta_e(0) = \theta_{e0}$ (Fig. 5.15), then $\dot{\theta}_e$ corresponding to this value of θ_e is negative. Hence, the phase error will start decreasing along the sinusoidal trajectory until it reaches the value θ_3, where equilibrium is attained. Hence, in steady state, the phase error is a constant θ_3. This means the loop is in frequency lock; that is, the VCO frequency is now ω_0, but there is a phase error of θ_3. Note, however, that if $|\omega_0 - \omega_c| > AK$, there are no

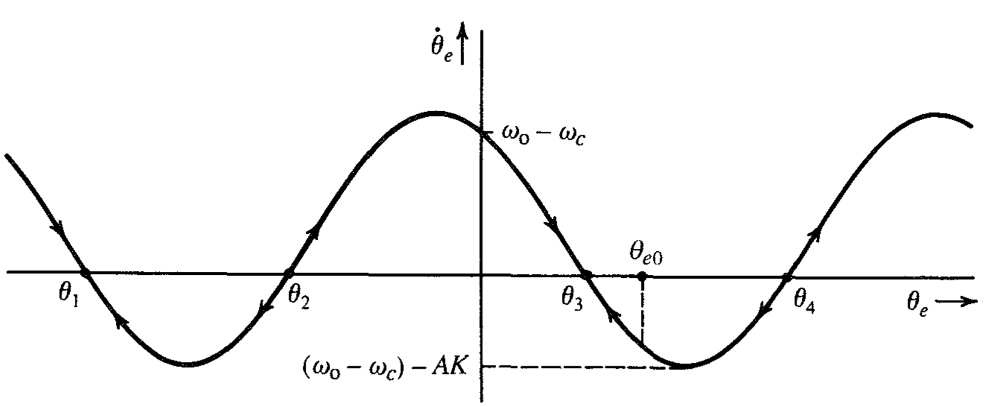


Figure 5.15 Trajectory of a first-order PLL.

---

* Actually $h(t) = 2B \operatorname{sinc}(2\pi Bt)$, where B is the bandwidth of the loop filter. This is a low-pass narrow-band filter, which suppresses the high-frequency signal centered at $2\omega_c$. This makes $H(s) = 1$ over a low-pass narrow band of B Hz.

equilibrium points in Fig. 5.15, the loop never achieves lock, and θ_e continues to move along the trajectory forever. Hence, this simple loop can achieve phase lock provided the incoming frequency ω_0 does not differ from the quiescent VCO frequency ω_c by more than AK.

In Fig. 5.15, several equilibrium points exist. Half of these points, however, are unstable equilibrium points, meaning that a slight perturbation in the system state will move the operating point farther away from these equilibrium points. Points θ_1 and θ_3 are stable points, because any small perturbation in the system state will tend to bring it back to these points. Consider, for example, the point θ_3. If the state is perturbed along the trajectory toward the right, $\dot{\theta}_e$ is negative, which tends to reduce θ_e and bring it back to θ_3. If the operating point is perturbed from θ_3 toward the left, $\dot{\theta}_e$ is positive, θ_e will tend to increase, and the operating point will return to θ_3. On the other hand, at point θ_2 if the point is perturbed toward the right, $\dot{\theta}_e$ is positive, and θ_e will increase until it reaches θ_3. Similarly, if at θ_2 the operating point is perturbed toward the left, $\dot{\theta}_e$ is negative, and θ_e will decrease until it reaches θ_1. Hence, θ_2 is an unstable equilibrium point. The slightest disturbance, such as noise, will dislocate it either to θ_1 or to θ_3. In a similar way, we can show that θ_4 is an unstable point and that θ_1 is a stable equilibrium point.

The equilibrium point θ_3 occurs where $\dot{\theta}_e = 0$. Hence, from Eq. (5.35),

$$\theta_3 = \sin^{-1} \frac{\omega_0 - \omega_c}{AK}$$

If $\theta_3 \ll \pi/2$, then

$$\theta_3 \simeq \frac{\omega_0 - \omega_c}{AK}$$

which agrees with our previous result of the small-error analysis [Eq. (5.31b)].

The first-order loop suffers from the fact that it has a constant phase error. Moreover, it can acquire frequency lock only if the incoming frequency and the VCO quiescent frequency differ by not more than AK rad/s. Higher order loops overcome these disadvantages, but they create a new problem of stability.^{10}

Another important class of detectors, the **FM demodulator with feedback (FMFB)**, uses feedback in the FM demodulator to narrow the bandwidth of the FM signal, which, in turn, reduces the noise power. This type of demodulator is discussed in Sec. 13.3.

5.5 INTERFERENCE IN ANGLE-MODULATED SYSTEMS

Let us consider the simple case of the interference of an unmodulated carrier $A \cos \omega_c t$ with another sinusoid $I \cos (\omega_c + \omega)t$. The received signal $r(t)$ is

$$\begin{aligned} r(t) &= A \cos \omega_c t + I \cos (\omega_c + \omega)t \\ &= (A + I \cos \omega t) \cos \omega_c t - I \sin \omega t \sin \omega_c t \\ &= E_r(t) \cos [\omega_c t + \psi_d(t)] \end{aligned}$$

where

$$\psi_d(t) = \tan^{-1} \frac{I \sin \omega t}{A + I \cos \omega t}$$

When the interfering signal is small in comparison to the carrier ($I \ll A$),

$$\psi_d(t) \simeq \frac{I}{A} \sin \omega t \tag{5.36}$$

The phase of $E_r(t) \cos [\omega_c t + \psi_d(t)]$ is $\psi_d(t)$, and its instantaneous frequency is $\omega_c + \dot{\psi}_d(t)$. If the signal $E_r(t) \cos [\omega_c t + \psi_d(t)]$ is applied to an ideal phase demodulator, the output $y_d(t)$ would be $\psi_d(t)$. Similarly, the output $y_d(t)$ of an ideal frequency demodulator would be $\dot{\psi}_d(t)$. Hence,

$$y_d(t) = \frac{I}{A} \sin \omega t \qquad \text{for PM} \tag{5.37}$$

$$y_d(t) = \frac{I\omega}{A} \cos \omega t \qquad \text{for FM} \tag{5.38}$$

Observe that in either case, the interference output is inversely proportional to the carrier amplitude A. Thus, the larger the carrier amplitude A, the smaller the interference effect. This behavior is very different from that in AM signals, where the interference output is independent of the carrier amplitude.* Hence, angle-modulated systems suppress weak interference ($I \ll A$) much better than do AM systems.

Because of the suppression of weak interference in FM, we observe what is known as the **capture effect**. For two transmitters with carrier-frequency separation less than the audio range, instead of getting interference, we observe that the stronger carrier effectively suppresses (captures) the weaker carrier. Subjective tests show that an interference level as low as 35 dB in the audio signals can cause objectionable effects. Hence, in AM, the interference level should be kept below 35 dB. On the other hand, for FM, because of the capture effect, the interference level need only be below 6 dB.

The interference amplitude (I/A for PM and $I\omega/A$ for FM) vs. ω at the receiver output is shown in Fig. 5.16. The interference amplitude is constant for all ω in PM but increases linearly with ω in FM.†

Interference Due to Channel Noise

The channel noise acts as interference in an angle-modulated signal. We shall consider the most common form of noise, white noise, which has a constant power spectral density. Such a noise may be considered as a sum of sinusoids of all frequencies in the band. All components have the same amplitudes (because of uniform PSD). This means I is constant for all ω, and the amplitude spectrum of the interference at the receiver output is as shown in Fig. 5.16. The interference amplitude spectrum is constant for PM, and increases linearly with ω for FM.

---

* For instance, an AM signal with an interfering sinusoid $I \cos (\omega_c + \omega)t$ is given by

$$\begin{aligned} r(t) &= [A + m(t)] \cos \omega_c t + I \cos (\omega_c + \omega)t \\ &= [A + m(t) + I \cos \omega t] \cos \omega_c t - I \sin \omega t \sin \omega_c t \end{aligned}$$

The envelope of this signal is

$$E(t) = \{[A + m(t) + I \cos \omega t]^2 + I^2 \sin^2 \omega t\}^{1/2} \approx A + m(t) + I \cos \omega t \qquad I \ll A$$

Thus the interference signal at the envelope detector output is $I \cos \omega t$, which is independent of the carrier amplitude A. We obtain the same result if synchronous demodulation is used. We come to a similar conclusion for AM-SC systems.

† The results in Eqs. (5.37) and (5.38) can be readily extended to more than one interfering sinusoid. The system behaves linearly for multiple interfering sinusoids provided their amplitudes are much smaller compared to the carrier amplitude.

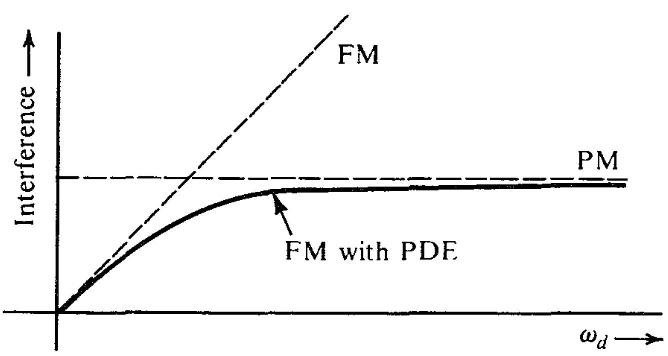


Figure 5.16 Effect of interference in PM, FM, and FM with preemphasis-deemphasis.

Preemphasis and Deemphasis in FM Broadcasting

Figure 5.16 shows that in FM, the interference (the noise) increases linearly with frequency, and the noise power in the receiver output is concentrated at higher frequencies. A glance at Fig. 4.19a shows that the PSD of an audio signal $m(t)$ is concentrated at lower frequencies below 2.1 kHz. Thus, the noise PSD is concentrated at higher frequencies, where $m(t)$ is weakest. This may seem like a disaster. But actually, in this very situation there is a hidden opportunity to reduce noise greatly. This process, shown in Fig. 5.17, works as follows: At the transmitter, the weaker high-frequency components (beyond 2.1 kHz) of the audio signal $m(t)$ are boosted before modulation by a **preemphasis** filter of transfer function $H_p(j\omega)$. At the receiver, the demodulator output is passed through a **deemphasis** filter of transfer function $H_d(\omega) = 1/H_p(j\omega)$. Thus, the deemphasis filter undoes the preemphasis by attenuating (deemphasizing) the higher frequency components (beyond 2.1 kHz), and thereby restores the original signal $m(t)$. The noise, however, enters at the channel, and therefore has not been preemphasized (boosted). However, it passes through the deemphasis filter, which attenuates its higher frequency components, where most of the noise power is concentrated (see Fig. 5.16). Thus, the process of preemphasis-deemphasis (PDE) leaves the desired signal untouched, but reduces the noise power considerably.

It may appear that we are gaining something for nothing. Not quite so! Boosting the higher frequency components of $m(t)$ increases its peak value m_p, which, in turn, increases $\Delta f = k_p m_p$. Thus, the preemphasis may seem to increase the transmission bandwidth. But the increase is minuscule because the (high-frequency) components that are boosted are so weak that even large amplification does not increase their absolute amplitude much. It is somewhat like a thousandfold increase in the salary of an unemployed person. A thousand times zero is still zero. Thus, preemphasis causes such a small increase in the signal power that the change in m_p is imperceptible, and we pay practically no price.

Preemphasis and Deemphasis Filters

Figure 5.16 indicates an approach to preemphasis. The FM has smaller interference than PM at lower frequencies, while the opposite is true at higher frequencies. If we can make our system behave like FM at lower frequencies and behave like PM at higher frequencies, we will have the best of both worlds. This is accomplished by a system used in commercial broadcasting (Fig. 5.17) with the preemphasis (before modulation) and deemphasis (after demodulation) filters $H_p(\omega)$ and $H_d(\omega)$ shown in Fig. 5.18. The frequency f_1 is 2.1 kHz, and f_2 is typically 30 kHz or more (well beyond audio range), so that f_2 does not even enter into the picture.

These filters can be realized by simple RC circuits (Fig. 5.18). The choice of $f_1 \doteq 2.1$ kHz was apparently made on an experimental basis. It was found that this choice of f_1 maintained the same peak amplitude m_p with or without preemphasis.[12] This satisfied the constraint of a fixed transmission bandwidth.

The preemphasis transfer function is

$$H_p(\omega) = K\frac{j\omega + \omega_1}{j\omega + \omega_2} \tag{5.39a}$$

where K, the gain, is set at a value of ω_2/ω_1. Thus,

$$H_p(\omega) = \left(\frac{\omega_2}{\omega_1}\right)\frac{j\omega + \omega_1}{j\omega + \omega_2} \tag{5.39b}$$

For $\omega \ll \omega_1$,

$$H_p(\omega) \simeq 1 \tag{5.39c}$$

For frequencies $\omega_1 \ll \omega \ll \omega_2$,

$$H_p(\omega) \simeq \frac{j\omega}{\omega_1} \tag{5.39d}$$

Thus, the preemphasizer acts as a differentiator at intermediate frequencies (2.1 to 15 kHz), which effectively makes the scheme PM over these frequencies. This means that FM with PDE

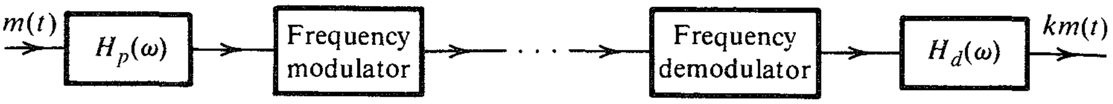

$m(t)$
$H_p(\omega)$
Frequency modulator
Frequency demodulator
$H_d(\omega)$
$km(t)$

Figure 5.17 Preemphasis-deemphasis in an FM system.

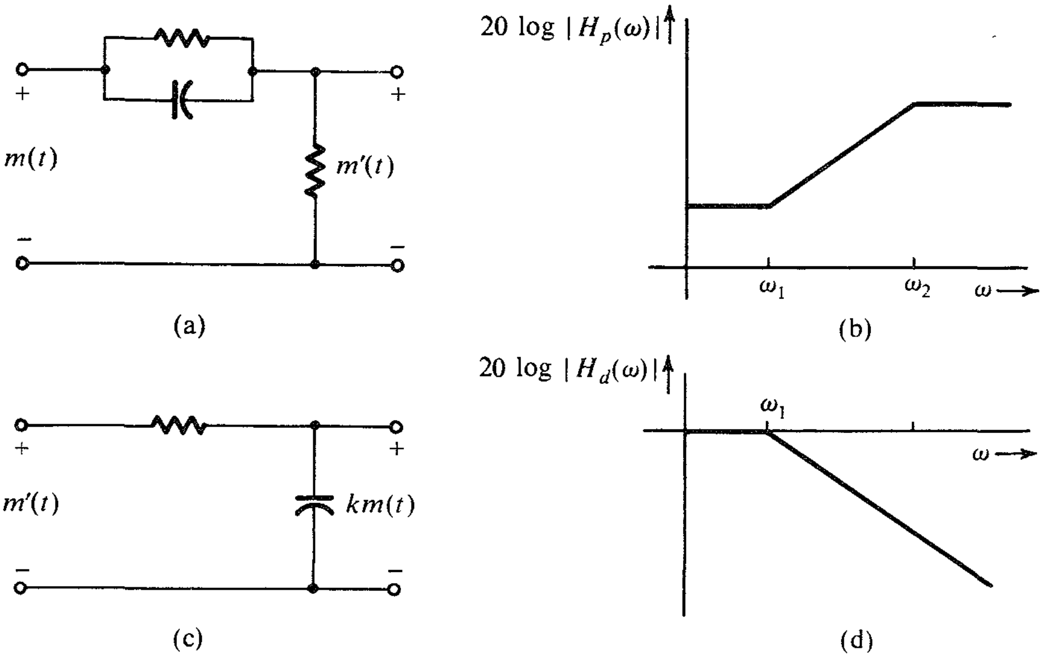

+
$m(t)$
−
+
$m'(t)$
−
(a)
20 log $|H_p(\omega)|$
ω_1
ω_2
ω
(b)
+
$m'(t)$
−
+
$km(t)$
−
(c)
20 log $|H_d(\omega)|$
ω_1
ω
(d)

Figure 5.18 (a) Preemphasis filter. (b) Its frequency response. (c) Deemphasis filter. (d) Its frequency response.

is FM over the modulating-signal frequency range of 0 to 2.1 kHz and is nearly PM over the range of 2.1 to 15 kHz as desired.

The deemphasis filter $H_d(\omega)$ is given by

$$H_d(\omega) = \frac{\omega_1}{j\omega + \omega_1}$$

Note that for $\omega \ll \omega_2$, $H_p(\omega) \simeq (j\omega + \omega_1)/\omega_1$. Hence, $H_p(\omega)H_d(\omega) \simeq 1$ over the baseband of 0 to 15 kHz.

Optimum PDE filters are discussed in Chapter 12. For historical and practical reasons, optimum PDE filters are not used in practice. It can be shown that the PDE enhances the SNR by 13.27 dB (a power ratio of 21.25).

The side benefit of PDE is improvement in the interference characteristics. Because the interference (from unwanted signals and the neighboring stations) enters after the transmitter stage, it undergoes only the deemphasis operation and not the boosting, or preemphasis. Hence, the interference amplitudes for frequencies beyond 2.1 kHz undergo attenuation that is roughly linear with frequency.

The PDE method of noise reduction is not limited just to FM broadcast. It is also used in audiotape recording and in (analog) phonograph recording, where the hissing noise is also concentrated at the high-frequency end. Sharp hissing sound is caused by irregularities in the recording material. The **Dolby noise reduction** systems for audiotapes operates on the same principle, although the Dolby-A system is somewhat more elaborate. In the Dolby-B and Dolby-C systems, the band is divided into two subbands (below and above 3 kHz instead of 2.1 kHz). In the Dolby-A system, designed for commercial use, the bands are divided into four subbands (below 80 Hz, between 80 Hz and 3 kHz, between 3 and 9 kHz, and above 9 kHz). The amount of preemphasis is optimized for each band.

We could also use PDE in AM broadcasting to improve the output SNR. In practice, however, this is not done for several reasons. First, the output noise amplitude in AM is constant with frequency, and does not increase linearly as in FM. Hence, the deemphasis does not yield such a dramatic improvement in AM as it does in FM. Second, introduction of PDE would necessitate modifications in receivers already in use. Third, increasing high-frequency component amplitudes (preemphasis) would increase interference with adjacent stations (no such problem arises in FM). Moreover, an increase in the deviation ratio (modulation index) at high frequencies would make detector design more difficult.

5.6 FM RECEIVER

The FCC has assigned a frequency range of 88 to 108 MHz for FM broadcasting, with a separation of 200 kHz between adjacent stations and a peak frequency deviation $\Delta f = 75$ kHz.

A monophonic FM receiver is identical to the superheterodyne AM receiver in Fig. 4.28, except that the intermediate frequency is 10.7 MHz and the envelope detector is replaced by a PLL or a frequency discriminator followed by a deemphasizer.

Earlier FM broadcasts were monophonic. Stereophonic FM broadcasting, in which two audio signals L (left microphone) and R (right microphone) are used for a more natural effect, was proposed later. The FCC ruled that the stereophonic system had to be compatible with the original monophonic system. This meant that the older monophonic receivers should be able to receive the signal $L + R$, and the total transmission bandwidth for the two signals (L

and R) should still be 200 kHz, with $\Delta f = 75$ kHz for the two combined signals. This would ensure that the older receivers could continue to receive monophonic as well as stereophonic broadcasts, although in the latter case the stereo effect would be absent.

A transmitter and a receiver for a stereo broadcast are shown in Fig. 5.19a and c. At the transmitter, the two signals L and R are added and subtracted to obtain $L+R$ and $L-R$. These signals are preemphasized. The preemphasized signal $(L-R)'$ DSB-SC modulates a carrier

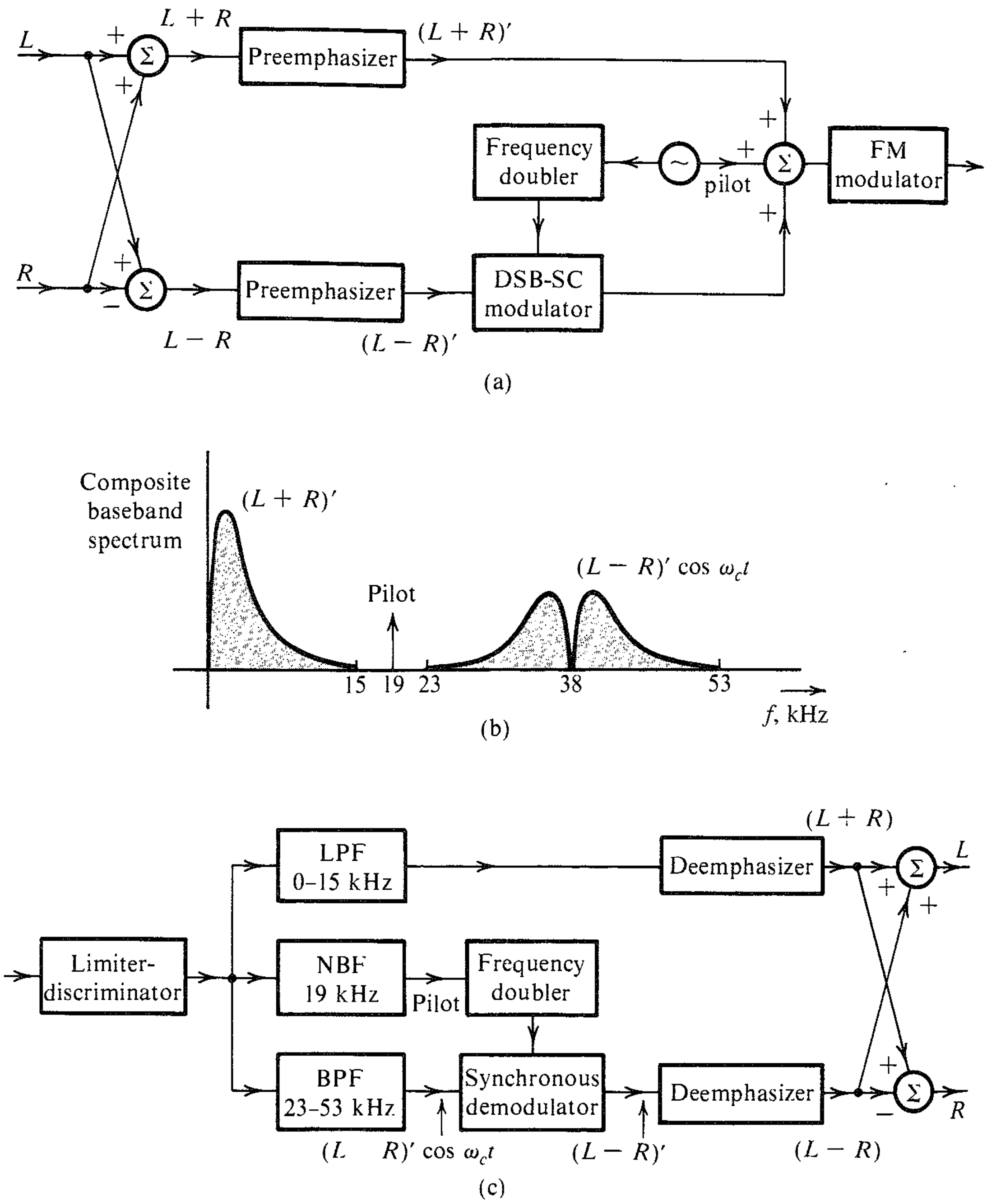

L
R
L + R
L − R
Σ
Preemphasizer
Preemphasizer
(L + R)'
(L − R)'
Frequency doubler
pilot
DSB-SC modulator
FM modulator
(a)
Composite baseband spectrum
(L + R)'
Pilot
(L − R)' cos ω_c t
15
19
23
38
53
f, kHz
(b)
Limiter-discriminator
LPF 0-15 kHz
NBF 19 kHz
Pilot
Frequency doubler
BPF 23-53 kHz
Synchronous demodulator
Deemphasizer
Deemphasizer
(L + R)
(L − R)
(L R)' cos ω_c t
(L − R)'
L
R
(c)

Figure 5.19 (a) FM stereo transmitter. (b) Spectrum of a baseband stereo signal. (c) FM stereo receiver.

of 38 kHz obtained by doubling the frequency of a 19-kHz signal that is used as a pilot. The signal $(L+R)'$ is used directly. All three signals (the third being the pilot) form a composite baseband signal $m(t)$ (Fig. 5.19b),

$$m(t) = (L+R)' + (L-R)' \cos \omega_c t + \alpha \cos \frac{\omega_c t}{2} \tag{5.40}$$

The reason for using a pilot of 19 kHz rather than 38 kHz is that it is easier to separate the pilot at 19 kHz, because there are no signal components within 4 kHz of that frequency.

The receiver operation (Fig. 5.19c) is self-explanatory. A monophonic receiver consists of only the upper branch of the stereo receiver and, hence, receives only $L+R$. This is of course the complete audio signal without the stereo effect. Hence, the system is compatible. The pilot is extracted, and (after doubling its frequency) it is used to demodulate coherently the signal $(L-R)' \cos \omega_c t$.

An interesting aspect of stereo transmission is that the peak amplitude of the composite signal $m(t)$ in Eq. (5.40) is practically the same as that of the monophonic signal (if we ignore the pilot), and, hence, Δf—which is proportional to the peak signal amplitude for stereophonic transmission—remains practically the same as for the monophonic case. This can be explained by the so-called **interleaving** effect as follows.

The L' and R' signals are very similar in general. Hence, we can assume their peak amplitudes to be equal to A_p. Under the worst possible conditions, L' and R' will reach their peaks at the same time, yielding [Eq. (5.40)]

$$|m(t)|_{\max} = 2A_p + \alpha$$

In the monophonic case, the peak amplitude of the baseband signal $(L+R)'$ is $2A_p$. Hence, the peak amplitudes in the two cases differ only by α, the pilot amplitude. To account for this, the peak sound amplitude in the stereo case is reduced to 90% of its full value. This amounts to a reduction in the signal power by a ratio of $(0.9)^2 = 0.81$, or 1 dB. Thus, the effective SNR is reduced by 1 dB because of the inclusion of the pilot.

REFERENCES

1. J. Carson, "Notes on the Theory of Modulation," *Proc. IRE*, vol. 10, pp. 57–64, Feb. 1922.
2. J. Carson, "Reduction of Atmospheric Disturbances," *Proc. IRE*, vol. 16, July 1928.
3. E. H. Armstrong, "A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation," *Proc. IRE*, vol. 24, pp. 689–740, May 1936.
4. L. Lessing, *Man of High Fidelity: Edwin Howard Armstrong*, J. B. Lippincott, Philadelphia, PA, 1956.
5. "A Revolution in Radio," *Fortune*, vol. 20, p. 116, Oct. 1939.
6. D. H. Sheingold, ed., *Nonlinear Circuits Handbook*, Analog Devices, Inc., Norwood, MA, 1974.
7. W. B. Davenport, Jr., "Signal-to-Noise Ratios in Bandpass Limiters," *J. Appl. Phys.*, vol. 24, pp. 720–727, June 1953.
8. H. L. Krauss, C. W. Bostian, and F. H. Raab, *Solid-State Radio Engineering*, Wiley, New York, 1980.
9. B. P. Lathi, *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, Carmichael, CA, 1998.
10. A. J. Viterbi, *Principles of Coherent Communication*, McGraw-Hill, New York, 1966.

11. W. C. Lindsey, *Synchronization Systems in Communication and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1972.
12. L. B. Arguimbau, and R. B. Adler, *Vacuum Tube Circuits and Transistors*, Wiley, New York, 1964, p. 466.

PROBLEMS

5.1-1 Sketch $\varphi_{\text{FM}}(t)$ and $\varphi_{\text{PM}}(t)$ for the modulating signal $m(t)$ shown in Fig. P5.1-1, given $\omega_c = 10^8$, $k_f = 10^5$, and $k_p = 25$.

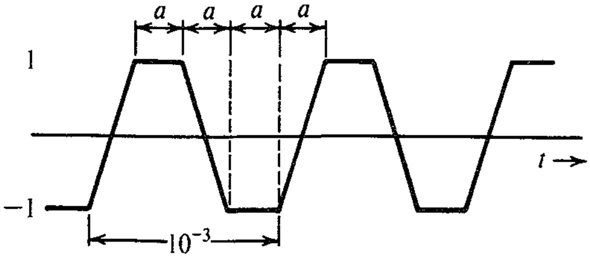


Figure P5.1-1

5.1-2 A baseband signal $m(t)$ is the periodic sawtooth signal shown in Fig. P5.1-2. Sketch $\varphi_{\text{FM}}(t)$ and $\varphi_{\text{PM}}(t)$ for this signal $m(t)$ if $\omega_c = 2\pi \times 10^6$, $k_f = 2000\pi$, and $k_p = \pi/2$. Explain why it is necessary to use $k_p < \pi$ in this case.

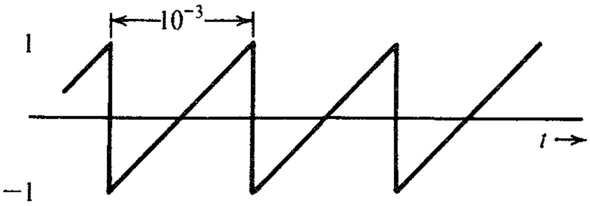


Figure P5.1-2

5.1-3 Over an interval $|t| \leq 1$, an angle modulated signal is given by

$$\varphi_{\text{EM}}(t) = 10 \cos 13,000t$$

It is known that the carrier frequency $\omega_c = 10,000$.

(a) If this were a PM signal with $k_p = 1000$, determine $m(t)$ over the interval $|t| \leq 1$.

(b) If this were an FM signal with $k_f = 1000$, determine $m(t)$ over the interval $|t| \leq 1$.

5.2-1 For a modulating signal

$$m(t) = 2\cos 100t + 18\cos 2000\pi t$$

(a) Write expressions (do not sketch) for $\varphi_{\text{PM}}(t)$ and $\varphi_{\text{FM}}(t)$ when $A = 10$, $\omega_c = 10^6$, $k_f = 1000\pi$, and $k_p = 1$. For determining $\varphi_{\text{FM}}(t)$, use the indefinite integral of $m(t)$, that is, take the value of the integral at $t = -\infty$ to be 0.

(b) Estimate the bandwidths of $\varphi_{\text{FM}}(t)$ and $\varphi_{\text{PM}}(t)$.

5.2-2 An angle-modulated signal with carrier frequency $\omega_c = 2\pi \times 10^6$ is described by the equation

$$\varphi_{\text{EM}}(t) = 10 \cos (\omega_c t + 0.1 \sin 2000\pi t)$$

(a) Find the power of the modulated signal.

(b) Find the frequency deviation Δf.

(c) Find the phase deviation $\Delta \phi$.

(d) Estimate the bandwidth of $\varphi_{\text{EM}}(t)$.

5.2-3 Repeat Prob. 5.2-2 if

$$\varphi_{\text{EM}}(t) = 5 \cos (\omega_c t + 20 \sin 1000\pi t + 10 \sin 2000\pi t)$$

5.2-4 Estimate the bandwidth for $\varphi_{\text{PM}}(t)$ and $\varphi_{\text{FM}}(t)$ in Prob. 5.1-1. Assume the bandwidth of $m(t)$ in Fig. P5.1-1 to be the third-harmonic frequency of $m(t)$.

5.2-5 Estimate the bandwidth of $\varphi_{\text{PM}}(t)$ and $\varphi_{\text{FM}}(t)$ in Prob. 5.1-2. Assume the bandwidth of $m(t)$ to be the fifth harmonic frequency of $m(t)$.

5.2-6 Given $m(t) = \sin 2000\pi t$, $k_f = 200,000\pi$, and $k_p = 10$.

(a) Estimate the bandwidths of $\varphi_{\text{FM}}(t)$ and $\varphi_{\text{PM}}(t)$.

(b) Repeat part **(a)** if the message signal amplitude is doubled.

(c) Repeat part **(a)** if the message signal frequency is doubled.

(d) Comment on the sensitivity of FM and PM bandwidths to the spectrum of $m(t)$.

5.2-7 Given $m(t) = e^{-t^2}$, $f_c = 10^4$ Hz, $k_f = 6000\pi$, and $k_p = 8000\pi$.

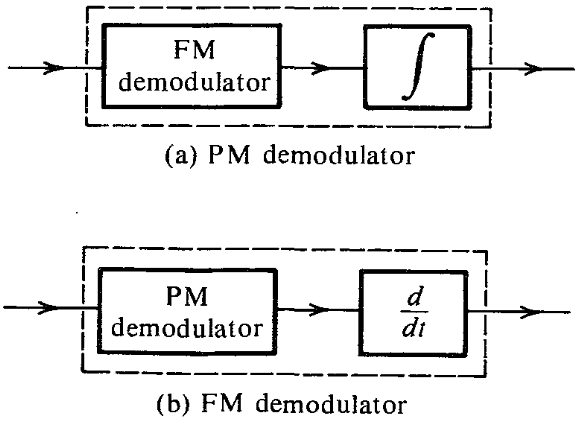
(a) Find Δf, the frequency deviation for FM and PM.

(b) Estimate the bandwidths of the FM and PM waves. *Hint*: Find $M(\omega)$ and observe the rapid decay of this spectrum. Its 3-dB bandwidth is even smaller than 1 Hz ($B \ll \Delta f$).

5.3-1 Design (only the block diagram) an Armstrong indirect FM modulator to generate an FM carrier with a carrier frequency of 98.1 MHz and $\Delta f = 75$ kHz. A narrow-band FM generator is available at a carrier frequency of 100 kHz and a frequency deviation $\Delta f = 10$ Hz. The stock room also has an oscillator with an adjustable frequency in the range of 10 to 11 MHz. There are also plenty of frequency doublers, triplers, and quintuplers.

5.3-2 Design (only the block diagram) an Armstrong indirect FM modulator to generate an FM carrier with a carrier frequency of 96 MHz and $\Delta f = 20$ kHz. A narrow-band FM generator with $f_c = 200$ kHz and adjustable Δf in the range of 9 to 10 Hz is available. The stock room also has an oscillator with adjustable frequency in the range of 9 to 10 MHz. There is a bandpass filter with any center frequency, and only frequency doublers are available.

5.4-1 Show that when $m(t)$ has no jump discontinuities, an FM demodulator followed by an integrator (Fig. P5.4-1a) acts as a PM demodulator, and a PM demodulator followed by a differentiator (Fig. P5.4-1b) serves as an FM demodulator even if $m(t)$ has jump discontinuities. *Hint*: For an input $A \cos [\omega_c t + \psi(t)]$, the output of an ideal FM demodulator is $\dot{\psi}(t)$ and that of an ideal PM demodulator is $\psi(t)$.


FM demodulator
∫
PM demodulator
d/dt

(a) PM demodulator

(b) FM demodulator

Figure P5.4-1

5.4-2 A periodic square wave $m(t)$ (Fig. P5.4-2a) frequency-modulates a carrier of frequency $f_c = 10$ kHz with $\Delta f = 1$ kHz. The carrier amplitude is A. The resulting FM signal is demodulated, as shown in Fig. P5.4-2b by the method discussed in Sec. 5.4 (Fig. 5.11). Sketch the waveforms at points b, c, d, and e.

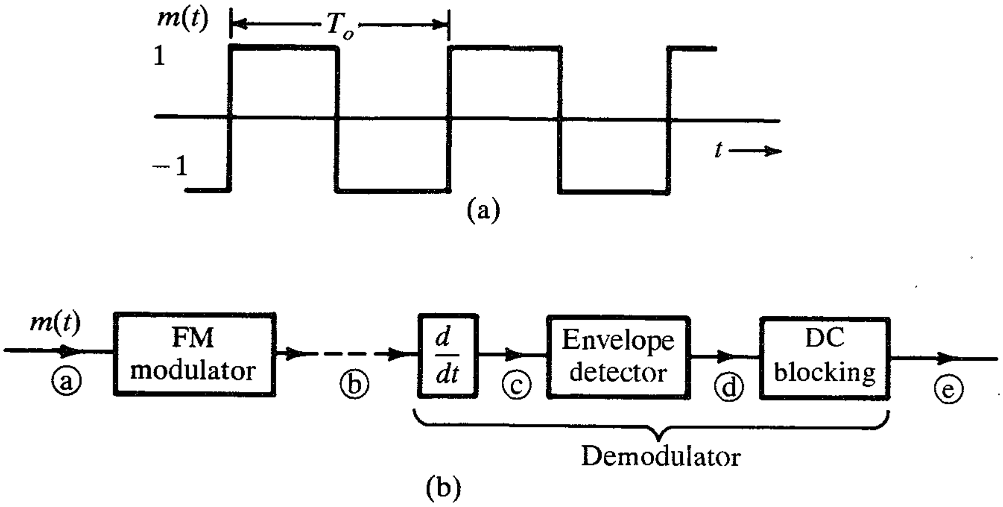

m(t)
T_o
1
−1
t
(a)
m(t)
ⓐ
FM modulator
ⓑ
d/dt
ⓒ
Envelope detector
ⓓ
DC blocking
ⓔ
Demodulator
(b)

Figure P5.4-2

5.4-3 Using small-error analysis, show that a first-order loop $[H(s) = 1]$ cannot track an incoming signal whose instantaneous frequency is varying linearly with time $[\theta_i(t) = kt^2]$. This signal can be tracked within a constant phase if $H(s) = (s + a)/s$. It can be tracked with a zero phase error if $H(s) = (s^2 + as + b)/s^2$.

6 SAMPLING AND PULSE CODE MODULATION

As seen in Chapter 1, analog signals can be digitized through sampling and quantization. The sampling rate must be sufficiently large so that the analog signal can be reconstructed from the samples with sufficient accuracy. The **sampling theorem**, which is the basis for determining the proper sampling rate for a given signal, has a deep significance in signal processing and communication theory.

6.1 SAMPLING THEOREM

We now show that a signal whose spectrum is band-limited to B Hz $[G(\omega) = 0 \text{ for } |\omega| > 2\pi B]$ can be reconstructed exactly (without any error) from its samples taken uniformly at a rate $R > 2B$ Hz (samples per second). In other words, the minimum sampling frequency is $f_s = 2B$ Hz.[*]

To prove the sampling theorem, consider a signal $g(t)$ (Fig. 6.1a) whose spectrum is band-limited to B Hz (Fig. 6.1b).[†] For convenience, spectra are shown as functions of ω as well as of f (Hz). Sampling $g(t)$ at a rate of f_s Hz (f_s samples per second) can be accomplished by multiplying $g(t)$ by an impulse train $\delta_{T_s}(t)$(Fig. 6.1c), consisting of unit impulses repeating periodically every T_s seconds, where $T_s = 1/f_s$. This results in the sampled signal $\overline{g}(t)$ shown in Fig. 6.1d. The sampled signal consists of impulses spaced every T_s seconds (the sampling interval). The nth impulse, located at $t = nT_s$, has a strength $g(nT_s)$, the value of $g(t)$ at $t = nT_s$. Thus,

$$\overline{g}(t) = g(t)\delta_{T_s}(t) = \sum_n g(nT_s)\delta(t - nT_s) \tag{6.1}$$

---

[*] The theorem stated here (and proved subsequently) applies to low-pass signals. A bandpass signal whose spectrum exists over a frequency band $f_c - B/2 < |f| < f_c + B/2$ has a bandwidth B Hz. Such a signal is uniquely determined by $2B$ samples per second. In general, the sampling scheme is a bit more complex in this case. It uses two interlaced sampling trains, each at a rate of B samples per second (known as second-order sampling). See, for example, the references 1, 2.

[†] The spectrum $G(\omega)$ in Fig. 6.1b is shown as real, for convenience. However, our arguments are valid for complex $G(\omega)$ as well.

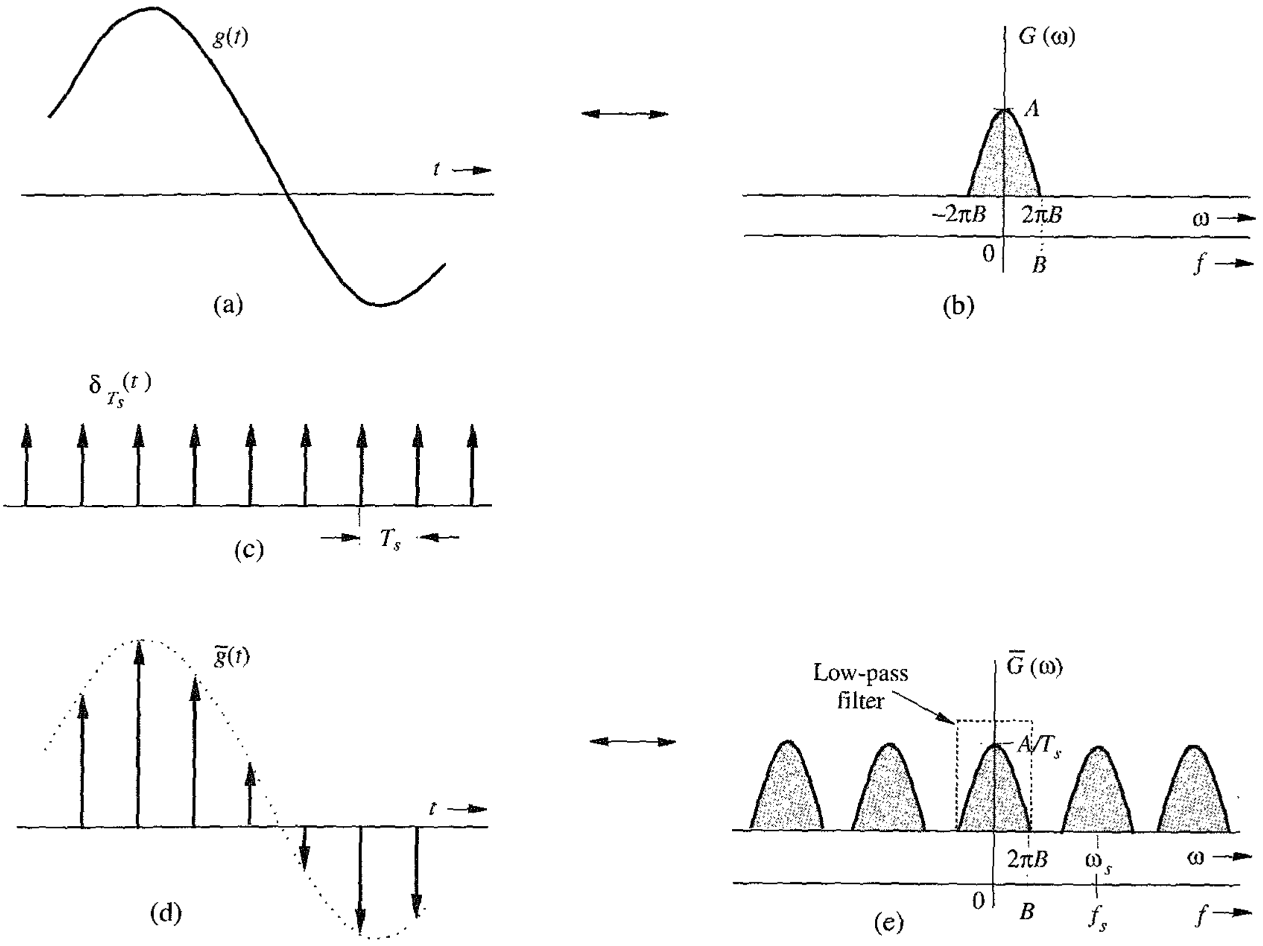


Figure 6.1 Sampled signal and its Fourier spectrum.

Because the impulse train $\delta_{T_s}(t)$ is a periodic signal of period T_s, it can be expressed as a Fourier series. The trigonometric Fourier series, already found in Example 2.9 [Eq. (2.77)], is

$$\delta_{T_s}(t) = \frac{1}{T_s}[1 + 2\cos\omega_s t + 2\cos 2\omega_s t + 2\cos 3\omega_s t + \cdots] \qquad \omega_s = \frac{2\pi}{T_s} = 2\pi f_s \tag{6.2}$$

Therefore,

$$\begin{aligned}\overline{g}(t) &= g(t)\delta_{T_s}(t)\\ &= \frac{1}{T_s}[g(t) + 2g(t)\cos\omega_s t + 2g(t)\cos 2\omega_s t + 2g(t)\cos 3\omega_s t + \cdots]\end{aligned} \tag{6.3}$$

To find $\overline{G}(\omega)$, the Fourier transform of $\overline{g}(t)$, we take the Fourier transform of the right-hand side of Eq. (6.3), term by term. The transform of the first term in the brackets is $G(\omega)$. The transform of the second term $2g(t)\cos\omega_s t$ is $G(\omega - \omega_s) + G(\omega + \omega_s)$ [see Eq. (3.35)]. This represents spectrum $G(\omega)$ shifted to ω_s and $-\omega_s$. Similarly, the transform of the third term $2g(t)\cos 2\omega_s t$ is $G(\omega - 2\omega_s) + G(\omega + 2\omega_s)$, which represents the spectrum $G(\omega)$ shifted to $2\omega_s$ and $-2\omega_s$, and so on to infinity. This means that the spectrum $\overline{G}(\omega)$ consists of $G(\omega$

repeating periodically with period $\omega_s = 2\pi/T_s$ rad/s, or $f_s = 1/T_s$ Hz, as shown in Fig. 6.1e. There is also a constant multiplier $1/T_s$ in Eq. (6.3). Therefore,

$$\overline{G}(\omega) = \frac{1}{T_s}\sum_{n=-\infty}^{\infty} G(\omega - n\omega_s) \tag{6.4}$$

If we are to reconstruct $g(t)$ from $\overline{g}(t)$, we should be able to recover $G(\omega)$ from $\overline{G}(\omega)$. This is possible if there is no overlap between successive cycles of $\overline{G}(\omega)$. Figure 6.1e shows that this requires

$$f_s > 2B \tag{6.5}$$

Also, the sampling interval $T_s = 1/f_s$. Therefore,

$$T_s < \frac{1}{2B} \tag{6.6}$$

Thus, as long as the sampling frequency f_s is greater than twice the signal bandwidth B (in hertz), $\overline{G}(\omega)$ will consist of nonoverlapping repetitions of $G(\omega)$. When this is true, Fig. 6.1e shows that $g(t)$ can be recovered from its samples $\overline{g}(t)$ by passing the sampled signal $\overline{g}(t)$ through an ideal low-pass filter of bandwidth B Hz. The minimum sampling rate $f_s = 2B$ required to recover $g(t)$ from its samples $\overline{g}(t)$ is called the **Nyquist rate** for $g(t)$, and the corresponding sampling interval $T_s = 1/2B$ is called the **Nyquist interval** for $g(t)$.*

6.1.1 Signal Reconstruction: The Interpolation Formula

The process of reconstructing a continuous-time signal $g(t)$ from its samples is also knows as **interpolation**. In Sec. 6.1, we saw that a signal $g(t)$ band-limited to B Hz can be reconstructed (interpolated) exactly from its samples. This is done by passing the sampled signal through an ideal low-pass filter of bandwidth B Hz. As seen from Eq. (6.3), the sampled signal contains a component $(1/T_s)g(t)$, and to recover $g(t)$ [or $G(\omega)$], the sampled signal must be passed through an ideal low-pass filter of bandwidth B Hz and gain T_s. Thus, the reconstruction (or interpolating) filter transfer function is

$$H(\omega) = T_s \operatorname{rect}\left(\frac{\omega}{4\pi B}\right) \tag{6.7}$$

The interpolation process here is expressed in the frequency domain as a filtering operation. Now, we shall examine this process from a different viewpoint, that of the time domain.

Let the signal interpolating (reconstruction) filter impulse response be $h(t)$. Thus, if we were to pass the sampled signal $\overline{g}(t)$ through this filter, its response would be $g(t)$. Let us now consider a very simple interpolating filter whose impulse response is rect (t/T_s), as shown in Fig. 6.2a. This is a gate pulse of unit height, centered at the origin, and of width T_s (the

---

* We have proved that the sampling rate $R > 2B$. However, if the spectrum $G(\omega)$ has no impulse (or its derivatives) at the highest frequency B, the signal can be recovered from its samples taken at a rate $R = 2B$ Hz (Nyquist rate). In case $G(\omega)$ contains an impulse at the highest frequency B, the rate R must be greater than $2B$ Hz. Such is the case when $g(t) = \sin 2\pi Bt$. This signal is band-limited to B Hz, but all of its samples are zero when taken at a rate $f_s = 2B$ (starting at $t = 0$), and $g(t)$ cannot be recovered from its Nyquist samples.

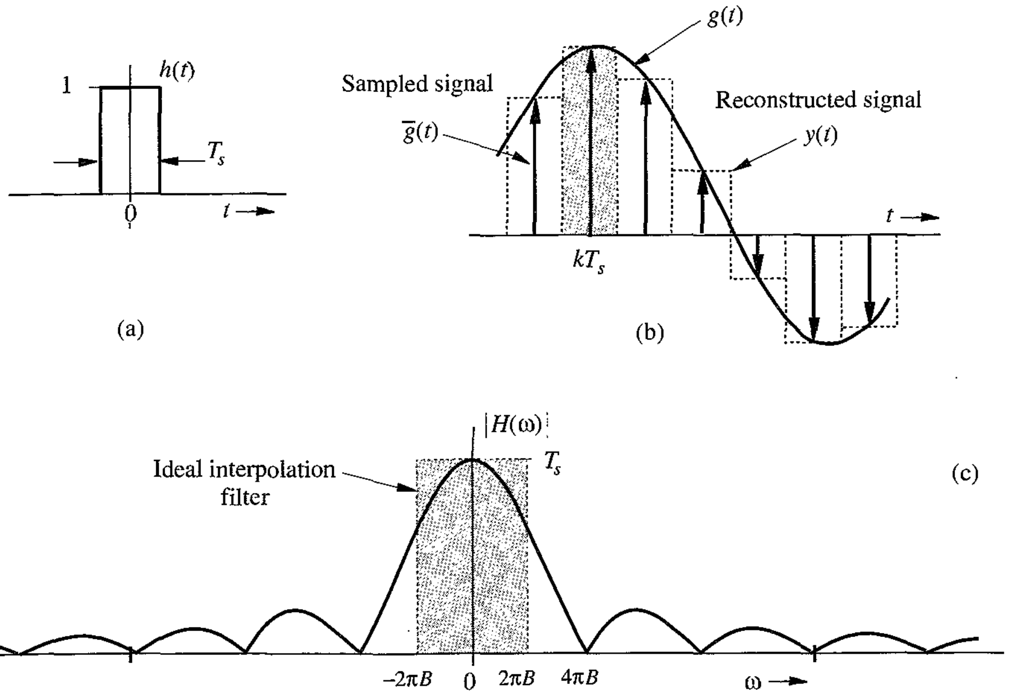


Figure 6.2 Simple interpolation using zero-order hold circuit.

sampling interval). Each sample in $\bar{g}(t)$, being an impulse, generates a gate pulse of the height equal to the strength of the sample. For instance, the kth sample is an impulse of strength $g(kT_s)$ located at $t = kT_s$, and can be expressed as $g(kT_s)\delta(t - kT_s)$. When this impulse passes through the filter, it generates an output $g(kT_s)\,\text{rect}\,(t/T_s)$. This is a gate pulse of height $g(kT_s)$, centered at $t = kT_s$ (shown shaded in Fig. 6.2b). Each sample in $\bar{g}(t)$ will generate a corresponding gate pulse resulting in an output

$$y(t) = \sum_k g(kT_s)\,\text{rect}\left(\frac{t}{T_s}\right)$$

The filter output is a staircase approximation of $g(t)$, shown dotted in Fig. 6.2b. This filter thus gives a crude form of interpolation.

The transfer function of this filter $H(\omega)$ is the Fourier transform of the impulse response $\text{rect}(t/T_s)$. Assuming the Nyquist sampling rate, that is, $T_s = 1/2B$,

$$h(t) = \text{rect}\left(\frac{t}{T_s}\right) = \text{rect}\,(2Bt)$$

and

$$H(\omega) = T_s\,\text{sinc}\left(\frac{\omega T_s}{2}\right) = \frac{1}{2B}\,\text{sinc}\left(\frac{\omega}{4B}\right) \tag{6.8}$$

The amplitude response $|H(\omega)|$ for this filter, shown in Fig. 6.2c, explains the reason for the crudeness of this interpolation. This filter, also known as the **zero-order hold** filter, is a poor approximation of the ideal low-pass filter (shown shaded in Fig. 6.2c) required for exact interpolation.*

---

* Figure 6.2a shows that the impulse response of this filter is noncausal, and this filter is not realizable. In practice we make it realizable by delaying the impulse response by $T_s/2$. This merely delays the output of the filter by $T_s/2$.

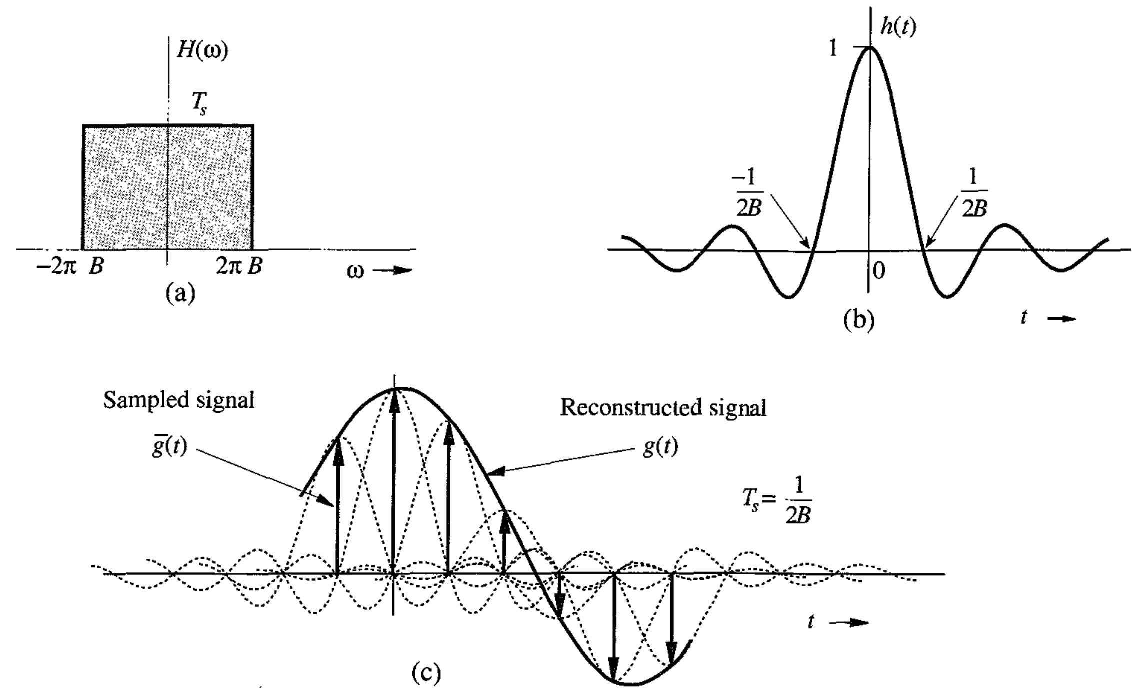


Figure 6.3 Ideal interpolation.

We can improve on the zero-order hold filter by using the **first-order hold** filter, which results in a linear interpolation instead of the staircase interpolation. The linear interpolator, whose impulse response is a triangle pulse $\Delta(t/2T_s)$, results in an interpolation in which successive sample tops are connected by straight-line segments (see Prob. 6.1-7).

The ideal interpolation filter transfer function found in Eq. (6.7) is shown in Fig. 6.3a. The impulse response of this filter, the inverse Fourier transform of $H(\omega)$, is

$$h(t) = 2BT_s \operatorname{sinc}(2\pi Bt) \tag{6.9a}$$

Assuming the Nyquist sampling rate, that is, $2BT_s = 1$, then

$$h(t) = \operatorname{sinc}(2\pi Bt) \tag{6.9b}$$

This $h(t)$ is shown in Fig. 6.3b. Observe the very interesting fact that $h(t) = 0$ at all Nyquist sampling instants ($t = \pm n/2B$) except at $t = 0$. When the sampled signal $\bar{g}(t)$ is applied at the input of this filter, the output is $g(t)$. Each sample in $\bar{g}(t)$, being an impulse, generates a sinc pulse of height equal to the strength of the sample, as shown in Fig. 6.3c. The process is identical to that shown in Fig. 6.2b, except that $h(t)$ is a sinc pulse instead of a gate pulse. Addition of the sinc pulses generated by all the samples results in $g(t)$. The kth sample of the input $\bar{g}(t)$ is the impulse $g(kT_s)\delta(t - kT_s)$; the filter output of this impulse is $g(kT_s)h(t - kT_s)$. Hence, the filter output to $\bar{g}(t)$, which is $g(t)$, can now be expressed as a sum,

$$\begin{aligned} g(t) &= \sum_k g(kT_s)h(t - kT_s) \\ &= \sum_k g(kT_s) \operatorname{sinc}[2\pi B(t - kT_s)] \qquad (6.10a) \\ &= \sum_k g(kT_s) \operatorname{sinc}(2\pi Bt - k\pi) \qquad (6.10b) \end{aligned}$$

Equation (6.10) is the **interpolation formula**, which yields values of $g(t)$ between samples as a weighted sum of all the sample values.

EXAMPLE 6.1 Find a signal $g(t)$ that is band-limited to B Hz and whose samples are

$$g(0) = 1 \quad \text{and} \quad g(\pm T_s) = g(\pm 2T_s) = g(\pm 3T_s) = \cdots = 0$$

where the sampling interval T_s is the Nyquist interval for $g(t)$, that is, $T_s = 1/2B$.

We use the interpolation formula (6.10b) to construct $g(t)$ from its samples. Since all but one of the Nyquist samples are zero, only one term (corresponding to $k = 0$) in the summation on the right-hand side of Eq. (6.10b) survives. Thus,

$$g(t) = \text{sinc}\,(2\pi Bt) \tag{6.11}$$

This signal is shown in Fig. 6.4. Observe that this is the only signal that has a bandwidth B Hz and with the sample values $g(0) = 1$ and $g(nT_s) = 0\,(n \neq 0)$. No other signal satisfies these conditions.

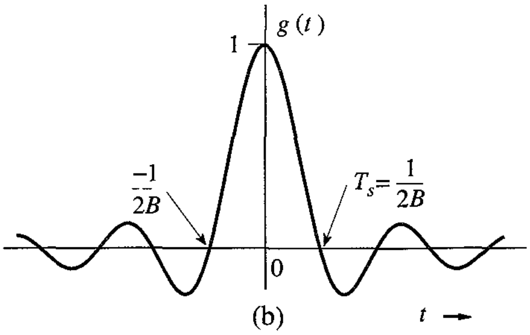


Figure 6.4 Signal reconstructed from the Nyquist samples in Example 6.1.

6.1.2 Practical Difficulties in Signal Reconstruction

If a signal is sampled at the Nyquist rate $f_s = 2B$ Hz, the spectrum $\overline{G}(\omega)$ consists of repetitions of $G(\omega)$ without any gap between successive cycles, as shown in Fig. 6.5a. To recover $g(t)$ from $\overline{g}(t)$, we need to pass the sampled signal $\overline{g}(t)$ through an ideal low-pass filter, shown dotted in Fig. 6.5a. As seen in Sec. 3.5, such a filter is unrealizable; it can be closely approximated only with infinite time delay in the response. This means that we can recover the signal $g(t)$ from its samples with infinite time delay. A practical solution to this problem is to sample the signal at a rate higher than the Nyquist rate ($f_s > 2B$ or $\omega_s > 4\pi B$). This yields $\overline{G}(\omega)$, consisting of repetitions of $G(\omega)$ with a finite band gap between successive cycles, as shown in Fig. 6.5b. We can now recover $G(\omega)$ from $\overline{G}(\omega)$ using a low-pass filter with a gradual cutoff characteristic, shown dotted in Fig. 6.5b. But even in this case, the filter gain is required to be zero beyond the first cycle of $G(\omega)$ (see Fig. 6.5b). By the Paley-Wiener criterion, it is impossible to realize even this filter. The only advantage in this case is that the required filter can be closely approximated with a smaller time delay. This shows that it is impossible in practice to recover a band-limited signal $g(t)$ exactly from its samples, even if the sampling rate is higher than the Nyquist rate. However, as the sampling rate increases, the recovered signal approaches the desired signal more closely.

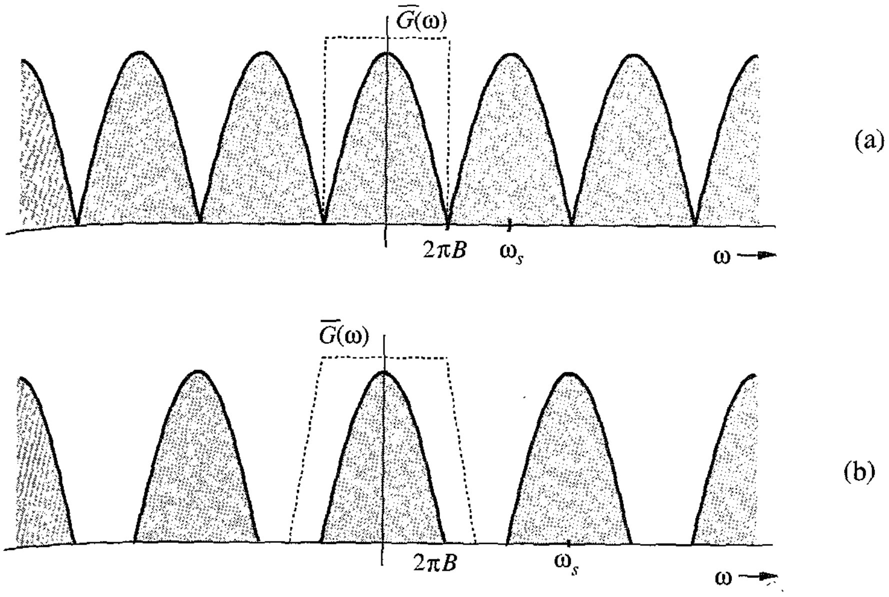


Figure 6.5 Spectra of a sampled signal. (a) At the Nyquist rate. (b) Above the Nyquist rate.

The Treachery of Aliasing

There is another fundamental practical difficulty in reconstructing a signal from its samples. The sampling theorem was proved on the assumption that the signal $g(t)$ is band-limited. **All practical signals are time-limited**, that is, they are of finite duration or width. It can be shown (see Prob. 6.1-8) that a signal cannot be time-limited and band-limited simultaneously. If a signal is time-limited, it cannot be band-limited, and vice versa (but it can be simultaneously non-time-limited and non-band-limited). This means that all practical signals, which are time-limited, are non-band-limited; they have infinite bandwidth, and the spectrum $\overline{G}(\omega)$ consists of overlapping cycles of $G(\omega)$ repeating every f_s Hz (the sampling frequency), as shown in Fig. 6.6. Because of infinite bandwidth in this case the spectral overlap is a constant feature, regardless of the sampling rate. Because of the overlapping tails, $\overline{G}(\omega)$ no longer has complete information about $G(\omega)$, and it is no longer possible, even theoretically, to recover $g(t)$ from the sampled signal $\overline{g}(t)$. If the sampled signal is passed through an ideal low-pass filter, the output is not $G(\omega)$ but a version of $G(\omega)$ distorted as a result of two separate causes:

1. The loss of the tail of $G(\omega)$ beyond $|f| > f_s/2$ Hz.
2. The reappearance of this tail inverted or folded onto the spectrum.

Note that the spectra cross at frequency $f_s/2 = 1/2T_s$ Hz. This frequency is called the **folding frequency**. The spectrum, therefore, folds onto itself at the folding frequency. For instance, a component of frequency $(f_s/2) + f_x$ shows up as or "impersonates" a component of lower frequency $(f_s/2) - f_x$ in the reconstructed signal. Thus, the components of frequencies above $f_s/2$ reappear as components of frequencies below $f_s/2$. This tail inversion, known as **spectral folding** or **aliasing**, is shown shaded in Fig. 6.6. In this process of aliasing, we are not only losing all the components of frequencies above $f_s/2$ Hz, but these very components reappear (aliased) as lower frequency components. This destroys the integrity of the lower frequency components also, as shown in Fig. 6.6.

The problem of aliasing is analogous to that of an army where a certain platoon has secretly defected to the enemy side. They are, however, nominally loyal to their army. The

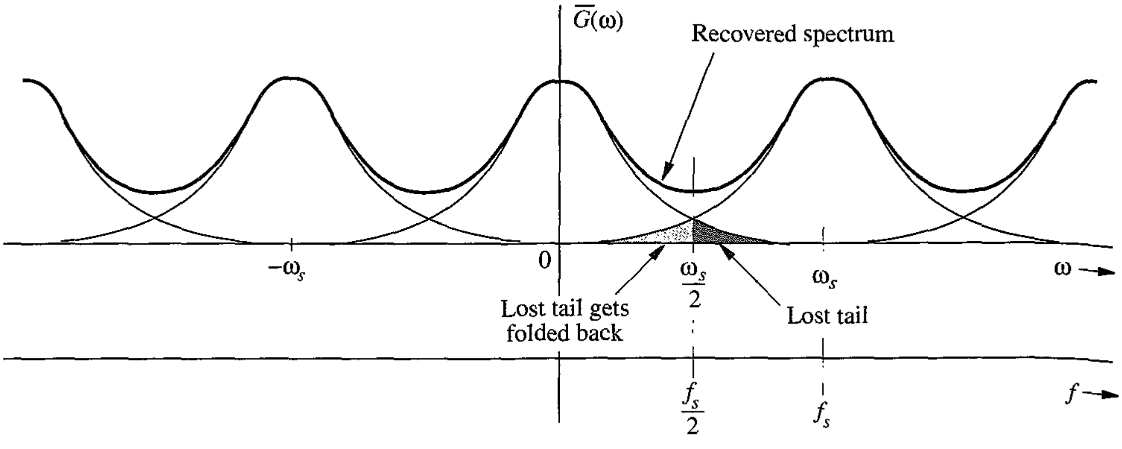


Figure 6.6 Aliasing effect.

army is in double jeopardy. First, they have lost this platoon as a fighting force. In addition, during actual fighting, the army will have to contend with the sabotage caused by the defectors, and will have to use another loyal platoon to neutralize the defectors. Thus, the army has lost two platoons in a nonproductive activity.

A Solution: The Antialiasing Filter

If you were the commander of the betrayed army, the solution to the problem would be obvious. As soon as the commander gets wind of the defection, he would incapacitate, by whatever means, the defecting platoon *before the fighting begins*. This way he loses only one (the defecting) platoon. This is a partial solution to the double jeopardy of betrayal. It partly rectifies the problem and reduces the losses to half.

We follow exactly the same procedure. The potential defectors are all the frequency components beyond $f_s/2 = 1/2T_s$ Hz. We should eliminate (suppress) these components from $g(t)$ *before sampling* $g(t)$. This way we lose only the components beyond the folding frequency $f_s/2$ Hz. These components now cannot reappear to corrupt the components with frequencies below the folding frequency. This suppression of higher frequencies can be accomplished by an ideal low-pass filter of bandwidth $f_s/2$ Hz. This filter is called the **antialiasing filter**. Note that the antialiasing operation must be performed *before the signal is sampled.*

The antialiasing filter, being an ideal filter, is unrealizable. In practice we use a steep cutoff filter, which leaves a sharply attenuated residual spectrum beyond the folding frequency $f_s/2$.

Practical Sampling

In proving the sampling theorem, we assumed ideal samples obtained by multiplying a signal $g(t)$ by an impulse train which is physically nonexistent. In practice, we multiply a signal $g(t)$ by a train of pulses of finite width, shown in Fig. 6.7b. The sampled signal is shown in Fig. 6.7c. We wonder whether it is possible to recover or reconstruct $g(t)$ from the sampled signal $\overline{g}(t)$ in Fig. 6.7c. Surprisingly, the answer is positive, provided that the sampling rate is not below the Nyquist rate. The signal $g(t)$ can be recovered by low-pass filtering $\overline{g}(t)$ as if it were sampled by impulse train.

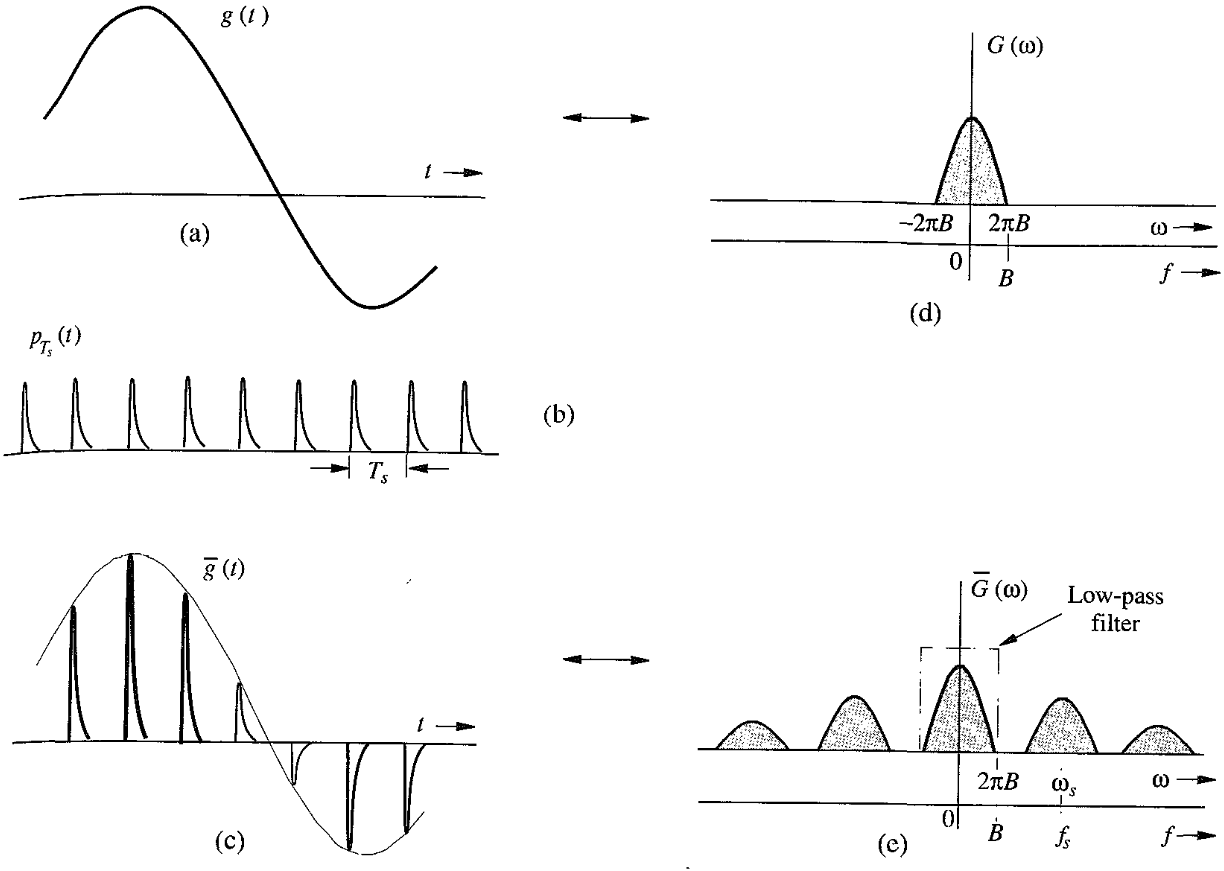


Figure 6.7 Sampled signal and its Fourier spectrum.

The plausibility of this result can be seen from the fact that to reconstruct $g(t)$, we need the knowledge of the Nyquist sample values. This information is available or built in the sampled signal $\bar{g}(t)$ in Fig. 6.7c because the kth sampled pulse strength is $g(kT_s)$. To prove the result analytically, we observe that the sampling pulse train $p_{T_s}(t)$ shown in Fig. 6.7b, being a periodic signal, can be expressed as a trigonometric Fourier series

$$p_{T_s}(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_s t + \theta_n) \qquad \omega_s = \frac{2\pi}{T_s}$$

and

$$\begin{aligned}\bar{g}(t) = g(t)p_{T_s}(t) &= g(t)\left[C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_s t + \theta_n)\right] \\ &= C_0 g(t) + \sum_{n=1}^{\infty} C_n g(t) \cos(n\omega_s t + \theta_n)\end{aligned}$$

The sampled signal $\bar{g}(t)$ consists of $C_0 g(t)$, $C_1 g(t)\cos(\omega_s t+\theta_1)$, $C_2 g(t)\cos(2\omega_s t+\theta_2)$, $\cdots$. Note that the first term $C_0 g(t)$ is the desired signal and all the other terms are modulated signals with spectra centered at $\pm\omega_s$, $\pm 2\omega_s$, $\pm\omega_s$, $\cdots$, as shown in Fig. 6.7e. Clearly the signal $g(t)$ can be recovered by low-pass filtering of $\bar{g}(t)$, provided that $\omega_s > 4\pi B$ (or $f_s > 2B$).

6.1.3 Maximum Information Rate: Two Pieces of Information per Second per Hertz

A knowledge of the maximum rate of information that can be transmitted over a channel of bandwidth B Hz is of fundamental importance in digital communication. We now derive one of the basic relationships in communication, which states that *a maximum of* $2B$ *independent pieces of information per second can be transmitted, errorfree, over a noiseless channel of bandwidth* B *Hz*. The result follows from the sampling theorem. Assuming no noise, a channel of bandwidth B Hz can transmit a signal of bandwidth B Hz errorfree. But a signal of bandwidth B can be reconstructed from its Nyquist samples, which are at a rate of $2B$ Hz. In other words, a signal of bandwidth B Hz can be completely specified by $2B$ independent pieces of information per second. Since the channel is capable of transmitting this signal errorfree, it follows that the channel should be able to transmit, errorfree, $2B$ independent pieces of information per second, and no more. In other words, *we can transmit errorfree at most two pieces of information per second per hertz bandwidth.* This is one of the fundamental relationships in communication theory.

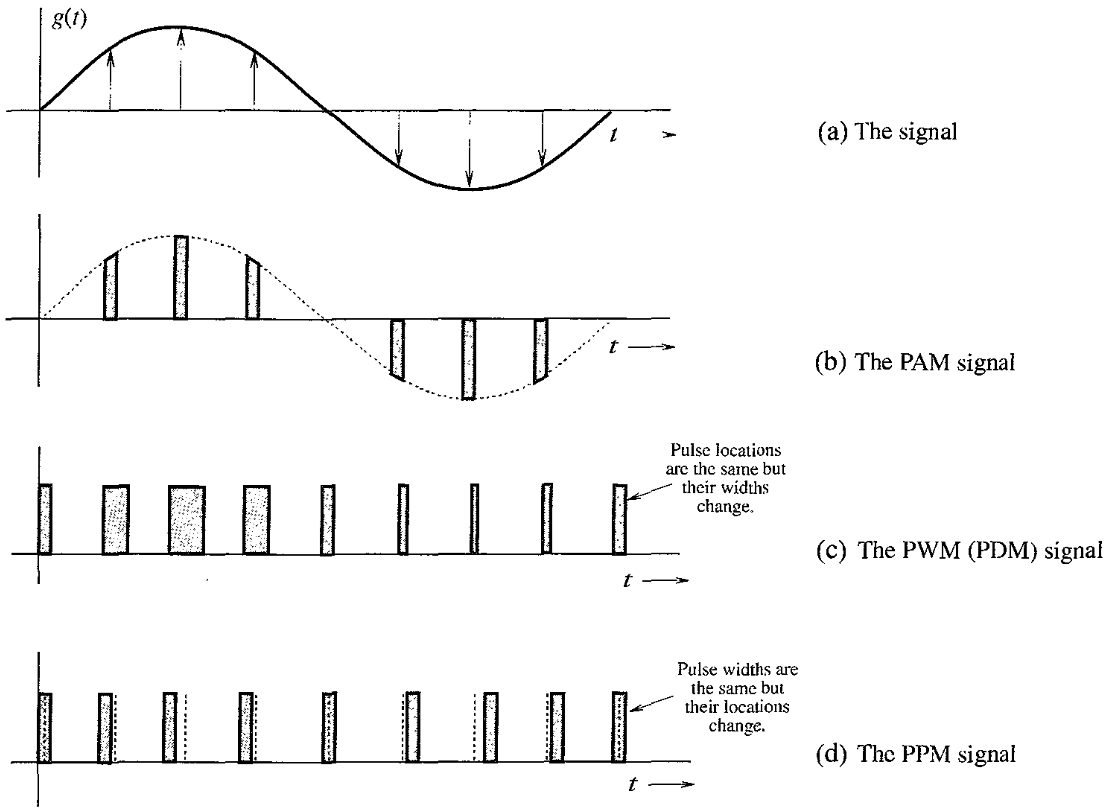
The Proof of the Pudding

To complete the proof, we now demonstrate a scheme which allows errorfree transmission of $2B$ independent pieces of information per second over a channel of bandwidth B Hz. Recall that a continuous-time signal $g(t)$ of bandwidth B Hz can be constructed from its Nyquist samples (which are at a rate of $2B$ Hz) using the interpolation formula (6.10). We can construct a signal from the given $2B$ pieces per second as the values of the Nyquist samples $g(kT_s)$ in Eq. (6.10). The resulting signal has a bandwidth B Hz and, therefore, can be transmitted errorfree over the channel of bandwidth B Hz. Moreover, the $2B$ pieces of information are readily obtained (errorfree) from this signal by taking its Nyquist samples.

This theoretical rate of communication assumes a noisefree channel. In practice, channel noise is unavoidable, and consequently, this rate will cause some detection errors. In Chapter 15, we shall determine the theoretical errorfree communication rate in the presence of noise.

6.1.4 Some Applications of the Sampling Theorem

The sampling theorem is very important in signal analysis, processing, and transmission because it allows us to replace a continuous-time signal by a discrete sequence of numbers. Processing a continuous-time signal is therefore equivalent to processing a discrete sequence of numbers. This leads us directly into the area of digital filtering. In the field of communication, the transmission of a continuous-time message reduces to the transmission of a sequence of numbers. This opens doors to many new techniques of communicating continuous-time signals by pulse trains. The continuous-time signal $g(t)$ is sampled, and sample values are used to modify certain parameters of a periodic pulse train. We may vary the amplitudes (Fig. 6.8b), widths (Fig. 6.8c), or positions (Fig. 6.8d) of the pulses in proportion to the sample values of the signal $g(t)$. Accordingly, we have **pulse-amplitude modulation** (PAM), **pulse-width modulation** (PWM), or **pulse-position modulation** (PPM). The most important form of pulse modulation today is **pulse-code modulation** (PCM), introduced in Sec. 1.2. In all these cases,



(a) The signal

(b) The PAM signal

Pulse locations are the same but their widths change.

(c) The PWM (PDM) signal

Pulse widths are the same but their locations change.

(d) The PPM signal

Figure 6.8 Pulse-modulated signals.

instead of transmitting $g(t)$, we transmit the corresponding pulse-modulated signal. At the receiver, we read the information of the pulse-modulated signal and reconstruct the analog signal $g(t)$.

One advantage of using pulse modulation is that it permits the simultaneous transmission of several signals on a time-sharing basis—**time-division multiplexing (TDM)**. Because a pulse-modulated signal occupies only a part of the channel time, we can transmit several pulse-modulated signals on the same channel by interweaving them. Figure 6.9 shows the TDM of two PAM signals. In this manner we can multiplex several signals on the same channel by reducing pulse widths.

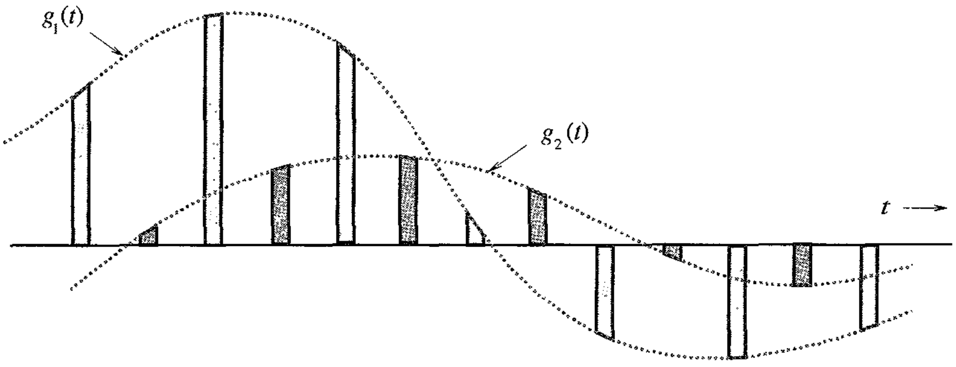


Figure 6.9 Time-division multiplexing of two signals.

Another method of transmitting several baseband signals simultaneously is frequency-division multiplexing (FDM), briefly discussed in Chapter 3. In FDM, various signals are multiplexed by sharing the channel bandwidth. The spectrum of each message is shifted to a specific band not occupied by any other signal. The information of various signals is located in nonoverlapping frequency bands of the channel. In a way, TDM and FDM are the duals of each other.

6.2 PULSE-CODE MODULATION (PCM)

PCM is the most useful and widely used of all the pulse modulations mentioned. Basically, PCM is a method of converting an analog signal into a digital signal (A/D conversion). An **analog** signal is characterized by the fact that its amplitude can take on any value over a continuous range. This means that it can take on an infinite number of values. On the other hand, **digital** signal amplitude can take on only a finite number of values. An analog signal can be converted into a digital signal by means of sampling and **quantizing**, that is, rounding off its value to one of the closest permissible numbers (or **quantized levels**), as shown in Fig. 6.10. The amplitudes of the analog signal $m(t)$ lie in the range $(-m_p, m_p)$, which is partitioned into L subintervals, each of magnitude $\Delta v = 2m_p/L$. Next, each sample amplitude is approximated by the midpoint value of the subinterval in which the sample falls (see Fig. 6.10 for $L = 16$). Each sample is now approximated to one of the L numbers. Thus, the signal is digitized, with quantized samples taking on any one of the L values. Such a signal is known as an **L-ary digital signal.**

From practical viewpoint, a binary digital signal (a signal that can take on only two values) is very desirable because of its simplicity, economy, and ease of engineering. We can convert an L-ary signal into a binary signal by using pulse coding. Figure 1.4b shows such a coding for the case of $L = 16$. This code, formed by binary representation of the 16 decimal digits from 0 to 15, is known as the **natural binary code (NBC)**. Other possible ways of assigning a binary code will be discussed later. Each of the 16 levels to be transmitted is

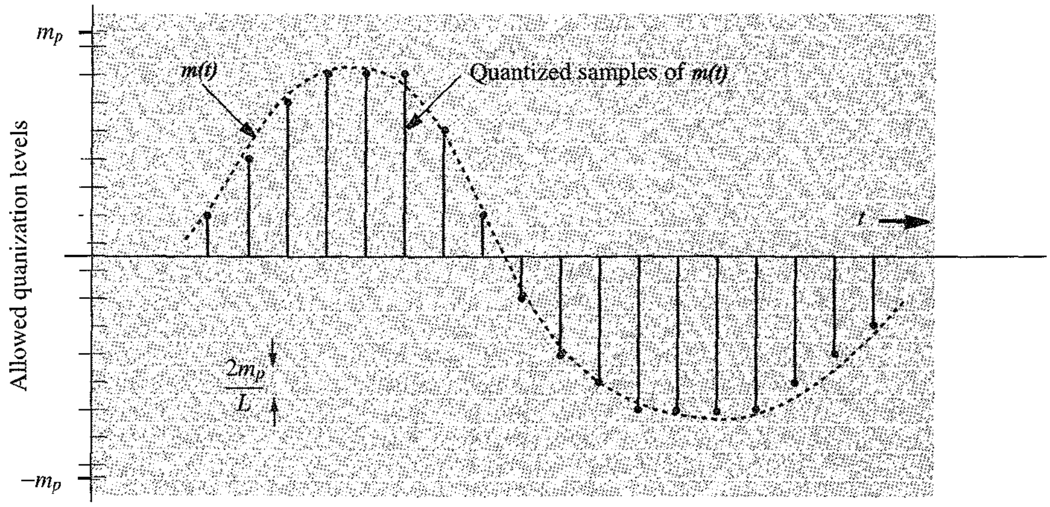


Figure 6.10 Quantization of a sampled analog signal.

assigned one binary code of four digits. The analog signal $m(t)$ is now converted to a (binary) digital signal. A **binary** digit is called a **bit** for convenience. This contraction of binary digit to bit has become an industry standard abbreviation and will be used throughout the book.

Thus, each sample in this example is encoded by four bits. To transmit this binary data, we need to assign a distinct pulse shape to each of the two bits. One possible way is to assign a negative pulse to a binary **0** and a positive pulse to a binary **1** (see Fig. 1.6) so that each sample is now transmitted by a group of four binary pulses (pulse code). The resulting signal is a binary signal.

The audio signal bandwidth is about 15 kHz, but subjective tests show that signal articulation (intelligibility) is not affected if all the components above 3400 Hz are suppressed.[3]* Since the objective in telephone communication is intelligibility rather than high fidelity, the components above 3400 Hz are eliminated by a low-pass filter. The resulting signal is then sampled at a rate of 8000 samples per second (8 kHz). This rate is intentionally kept higher than the Nyquist sampling rate of 6.8 kHz to avoid unrealizable filters required for signal reconstruction. Each sample is finally quantized into 256 levels ($L = 256$), which requires a group of eight binary pulses to encode each sample ($2^8 = 256$). Thus, a telephone signal requires $8 \times 8000 = 64000$ binary pulses per second.

The compact disc (CD) is a recent application of PCM. This is a high-fidelity situation requiring the audio signal bandwidth to be 15 kHz. Although the Nyquist sampling rate is only 30 kHz, the actual sampling rate of 44.1 kHz is used for the reason mentioned earlier. The signal is quantized into a rather large number of levels ($L = 65,536$) to reduce the quantizing error. The binary-coded samples are now recorded on the CD.

Advantages of Digital Communication

Some of the advantages of digital communication over analog communication are listed below:

1. Digital communication is more rugged than analog communication because it can withstand channel noise and distortion much better as long as the noise and the distortion are within limits, as shown in Fig. 1.3. Such is not the case with analog messages. Any distortion or noise, no matter how small, will distort the received signal.
2. The greatest advantage of digital communication over analog communication, however, is the viability of regenerative repeaters in the former. In an analog communication system, a message signal, as it travels along the channel (transmission path), grows progressively weaker, whereas the channel noise and the signal distortion, being cumulative, become progressively stronger. Ultimately the signal, overwhelmed by noise and distortion, is mutilated. Amplification is of little help because it enhances the signal and the noise in the same proportion. Consequently, the distance over which an analog message can be transmitted is limited by the transmitted power. If a transmission path is long enough, the channel distortion and noise will accumulate sufficiently to overwhelm even a digital signal. The trick is to set up repeater stations along the transmission path at distances short enough to be able to detect signal pulses before the noise and distortion have a chance to accumulate sufficiently. At each repeater station the pulses are detected, and new, clean pulses are transmitted to the next repeater station, which, in turn, duplicates the same process. If the noise and distortion are within limits (which is possible because of the

---

* Components below 300 Hz may also be suppressed without affecting the articulation.

closely spaced repeaters), pulses can be detected correctly.* This way the digital messages can be transmitted over longer distances with greater reliability. In contrast, analog messages cannot be cleaned up periodically, and the transmission is therefore less reliable. The most significant error in PCM comes from quantizing. This error can be reduced as much as desired by increasing the number of quantizing levels, the price of which is paid in an increased bandwidth of the transmission medium (channel).

3. Digital hardware implementation is flexible and permits the use of microprocessors, miniprocessors, digital switching, and large-scale integrated circuits.
4. Digital signals can be coded to yield extremely low error rates and high fidelity as well as privacy.
5. It is easier and more efficient to multiplex several digital signals.
6. Digital communication is inherently more efficient than analog in realizing the exchange of SNR for bandwidth.
7. Digital signal storage is relatively easy and inexpensive. It also has the ability to search and select information from distant electronic storehouses.
8. Reproduction with digital messages is extremely reliable without deterioration. Analog messages such as photocopies and films, for example, lose quality at each successive stage of reproduction, and have to be transported physically from one distant place to another, often at relatively high cost.
9. The cost of digital hardware continues to halve every two or three years, while performance or capacity doubles over the same time period. And there is no end in sight yet to this breathtaking and relentless exponential progress in digital technology. In recent years we have seen the compact disc—a digital device—bury the analog long-playing record; newspapers transmit photographs in scanned digital form; and more recently the shift in the United States toward a digital standard for high-definition television as opposed to the analog standard embraced by Japan and Europe. In contrast, analog technologies such as paper, video, sound, and film do not decline rapidly in cost. If anything, they become more expensive with time. For these and other reasons, it is only a matter of time before cost/performance curves cross, and digital technologies come to dominate in any given area of communication or storage technologies.

A Historical Note

Gottfried Wilhelm Leibnitz (1646-1716) was the first mathematician to work out systematically the binary representation (using **1**'s and **0**'s) for any number. He felt a spiritual significance in this discovery, reasoning that **1**, representing unity, was clearly a symbol for God, while **0** represented the nothingness. Therefore, if all numbers can be represented merely by the use of **1** and **0**, surely this proves that God created the universe out of nothing!

6.2.1 Quantizing

As mentioned earlier, digital signals come from a variety of sources. Some sources such as computers are inherently digital. Some sources are analog, but are converted into digital form by a variety of techniques such as PCM and delta modulation (DM), which will now be analyzed. The rest of this section provides quantitative discussion of PCM and its various aspects, such as quantizing, encoding, synchronizing, the required transmission bandwidth , the SNR, and so on.

---

* The error in pulse detection can be made negligible.

For quantizing, we limit the amplitude of the message signal $m(t)$ to the range $(-m_p, m_p)$, as shown in Fig. 6.10. Note that m_p is not necessarily the peak amplitude of $m(t)$. The amplitudes of $m(t)$ beyond $\pm m_p$ are chopped off. Thus, m_p is not a parameter of the signal $m(t)$, but is a constant of the quantizer. The amplitude range $(-m_p, m_p)$ is divided into L uniformly spaced intervals, each of width $\Delta v = 2m_p/L$. A sample value is approximated by the midpoint of the interval in which it lies (Fig. 6.10). The quantized samples are coded and transmitted as binary pulses. At the receiver some pulses will be detected incorrectly. Hence, there are two sources of error in this scheme: **quantization error** and **pulse detection error**. In almost all practical schemes, the pulse detection error is quite small compared to the quantization error and can be ignored. In the present analysis, therefore, we shall assume that the error in the received signal is caused exclusively by quantization. A general analysis that includes both types of errors is given in Sec. 12.4.

If $m(kT_s)$ is the kth sample of the signal $m(t)$, and if $\hat{m}(kT_s)$ is the corresponding quantized sample, then from the interpolation formula in Eq. (6.10),

$$m(t) = \sum_k m(kT_s) \operatorname{sinc}\,(2\pi Bt - k\pi)$$

and

$$\hat{m}(t) = \sum_k \hat{m}(kT_s) \operatorname{sinc}\,(2\pi Bt - k\pi)$$

where $\hat{m}(t)$ is the signal reconstructed from quantized samples. The distortion component $q(t)$ in the reconstructed signal is $q(t) = \hat{m}(t) - m(t)$. Thus,

$$\begin{aligned} q(t) &= \sum_k [\hat{m}(kT_s) - m(kT_s)] \operatorname{sinc}\,(2\pi Bt - k\pi) \\ &= \sum_k q(kT_s) \operatorname{sinc}\,(2\pi Bt - k\pi) \end{aligned}$$

where $q(kT_s)$ is the quantization error in the kth sample. The signal $q(t)$ is the undesired signal, and, hence, acts as noise, known as **quantization noise**. To calculate the power, or the mean square value of $q(t)$, we have

$$\begin{aligned} \widetilde{q^2(t)} &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} q^2(t)\,dt \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \left[\sum_k q(kT_s) \operatorname{sinc}\,(2\pi Bt - k\pi)\right]^2 dt \end{aligned} \tag{6.12a}$$

We can show that (see Prob. 3.7-4) the signals sinc $(2\pi Bt - m\pi)$ and sinc $(2\pi Bt - n\pi)$ are orthogonal, that is,

$$\int_{-\infty}^{\infty} \operatorname{sinc}\,(2\pi Bt - m\pi) \operatorname{sinc}\,(2\pi Bt - n\pi)\,dt = \begin{cases} 0 & m \neq n \\ \frac{1}{2B} & m = n \end{cases} \tag{6.12b}$$

Because of this result, the integrals of the cross-product terms on the right-hand side of Eq. (6.12a) vanish, and we obtain

$$\widetilde{q^2(t)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \sum_k q^2(kT_s)\, \mathrm{sinc}^2\,(2\pi Bt - k\pi)\, dt$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_k q^2(kT_s) \int_{-T/2}^{T/2} \mathrm{sinc}^2\,(2\pi Bt - k\pi)\, dt$$

From the orthogonality relationship (6.12b), it follows that

$$\widetilde{q^2(t)} = \lim_{T\to\infty} \frac{1}{2BT} \sum_k q^2(kT_s) \tag{6.13}$$

Because the sampling rate is $2B$, the total number of samples over the averaging interval T is $2BT$. Hence, the right-hand side of Eq. (6.13) represents the average, or the mean of the square of the quantization error. The quantum levels are separated by $\Delta v = 2m_p/L$. Since a sample value is approximated by the midpoint of the subinterval (of height Δv) in which the sample falls, the maximum quantization error is $\pm\Delta v/2$. Thus, the quantization error lies in the range $(-\Delta v/2,\ \Delta v/2)$, where

$$\Delta v = \frac{2m_p}{L} \tag{6.14}$$

Assuming that the error is equally likely to lie anywhere in the range $(-\Delta v/2,\ \Delta v/2)$, the mean square quantizing error $\widetilde{q^2}$ is given by*

$$\widetilde{q^2} = \frac{1}{\Delta v} \int_{-\Delta v/2}^{\Delta v/2} q^2\, dq$$

$$= \frac{(\Delta v)^2}{12} \tag{6.15a}$$

$$= \frac{m_p^2}{3L^2} \tag{6.15b}$$

Because $\widetilde{q^2(t)}$ is the mean square value or power of the quantization noise, we shall denote it by N_q,

$$N_q = \widetilde{q^2(t)} = \frac{m_p^2}{3L^2}$$

Assuming that the pulse detection error at the receiver is negligible, the reconstructed signal $\hat{m}(t)$ at the receiver output is

$$\hat{m}(t) = m(t) + q(t)$$

---

* Those who are familiar with the theory of probability can derive this result directly by noting that the probability density of the quantization error q is $1/(2m_p/L) = L/2m_p$ over the range $|q| \le m_p/L$ and is zero elsewhere. Hence,

$$\overline{q^2} = \int_{-m_p/L}^{m_p/L} q^2 p(q)\, dq = \int_{-m_p/L}^{m_p/L} \frac{L}{2m_p} q^2\, dq = \frac{m_p^2}{3L^2}$$

The desired signal at the output is $m(t)$, and the (quantization) noise is $q(t)$. Since the power of the message signal $m(t)$ is $\widetilde{m^2(t)}$, then

$$S_o = \widetilde{m^2(t)}$$

$$N_o = N_q = \frac{m_p^2}{3L^2}$$

and

$$\frac{S_o}{N_o} = 3L^2\frac{\widetilde{m^2(t)}}{m_p^2} \tag{6.16}$$

In this equation m_p is the peak amplitude value that a quantizer accept, and is therefore a constant of the quantizer. This means S_o/N_o, the SNR, is a linear function of the message signal power $\widetilde{m^2(t)}$ (see Fig. 6.13 with $\mu = 0$).

6.2.2 Principle of Progressive Taxation: Nonuniform Quantization

Recall that S_o/N_o, the SNR, is an indication of the quality of the received signal. Ideally we would like to have a constant SNR (the same quality) for all values of the message signal power $\widetilde{m^2(t)}$. Unfortunately, the SNR is directly proportional to the signal power $\widetilde{m^2(t)}$, which varies from talker to talker by as much as 40 dB (a power ratio of 10^4). The signal power can also vary because of the different lengths of the connecting circuits. This means the SNR in Eq. (6.16) can vary widely, depending on the talker and the length of the circuit. Even for the same talker, the quality of the received signal will deteriorate markedly when the person speaks softly. Statistically, it is found that smaller amplitudes predominate in speech and larger amplitudes are much less frequent. This means the SNR will be low most of the time.

The root of this difficulty lies in the fact that the quantizing steps are of uniform value $\Delta v = 2m_p/L$. The quantization noise $N_q = (\Delta v)^2/12$ [Eq. (6.15b)] is directly proportional to the square of the step size. The problem can be solved by using smaller steps for smaller amplitudes (nonuniform quantizing), as shown in Fig. 6.11a. The same result is obtained by first compressing signal samples and then using a uniform quantization. The input-output characteristics of a compressor are shown in Fig. 6.11b. The horizontal axis is the normalized input signal (i.e., the input signal amplitude m divided by the signal peak value m_p). The vertical axis is the output signal y. The compressor maps input signal increments Δm into larger increments Δy for small input signals, and vice versa for large input signals. Hence, a given interval Δm contains a larger number of steps (or smaller step size) when m is small. The quantization noise is smaller for smaller input signal power. An approximately logarithmic compression characteristic yields a quantization noise nearly proportional to the signal power $\widetilde{m^2(t)}$, thus making the SNR practically independent of the input signal power over a large dynamic range[4, 5] (see Fig. 6.13). This approach of equalizing the SNR appears similar to the use of progressive income tax to equalize incomes. The loud talkers and stronger signals are penalized with higher noise steps Δv in order to compensate the soft talkers and weaker signals.

Among several choices, two compression laws have been accepted as desirable standards by the CCITT: the μ-law used in North America and Japan, and the A-law used in Europe and the rest of the world and international routes. Both the μ-law and the A-law curves have odd symmetry about the vertical axis. The μ-law (for positive amplitudes) is given by

$$y = \frac{1}{\ln(1+\mu)} \ln\left(1 + \frac{\mu m}{m_p}\right) \qquad 0 \leq \frac{m}{m_p} \leq 1 \tag{6.17a}$$

The A-law (for positive amplitudes) is

$$y = \begin{cases} \dfrac{A}{1+\ln A}\left(\dfrac{m}{m_p}\right) & 0 \leq \dfrac{m}{m_p} \leq \dfrac{1}{A} \\[2ex] \dfrac{1}{1+\ln A}\left(1 + \ln\dfrac{Am}{m_p}\right) & \dfrac{1}{A} \leq \dfrac{m}{m_p} \leq 1 \end{cases} \tag{6.17b}$$

These characteristics are shown in Fig. 6.12.

The compression parameter μ (or A) determines the degree of compression. To obtain a nearly constant S_o/N_o over an input-signal-power dynamic range of 40 dB, μ should be greater than 100. Early North American channel banks and other digital terminals used a value of $\mu = 100$, which yielded the best results for 7-bit (128-level) encoding. An optimum value of $\mu = 255$ has been used for all North American 8-bit (256-level) digital terminals, and the earlier value of μ is now almost extinct. For the A-law, a value of $A = 87.6$ gives comparable results and has been standardized by the CCITT.

The compressed samples must be restored to their original values at the receiver by using an expander with a characteristic complementary to that of the compressor. The compressor and the expander together are called the **compandor**. Generally speaking, compression of a signal increases its bandwidth. But in PCM, we are compressing not the signal $m(t)$ but its samples. Because the number of samples does not change, the problem of bandwidth increase does not arise here.

It is shown in Sec. 12.4 that when a μ-law compandor is used, the output SNR is

$$\frac{S_o}{N_o} \simeq \frac{3L^2}{[\ln(1+\mu)]^2} \qquad \mu^2 \gg \frac{m_p^2}{\widetilde{m^2(t)}} \tag{6.18}$$

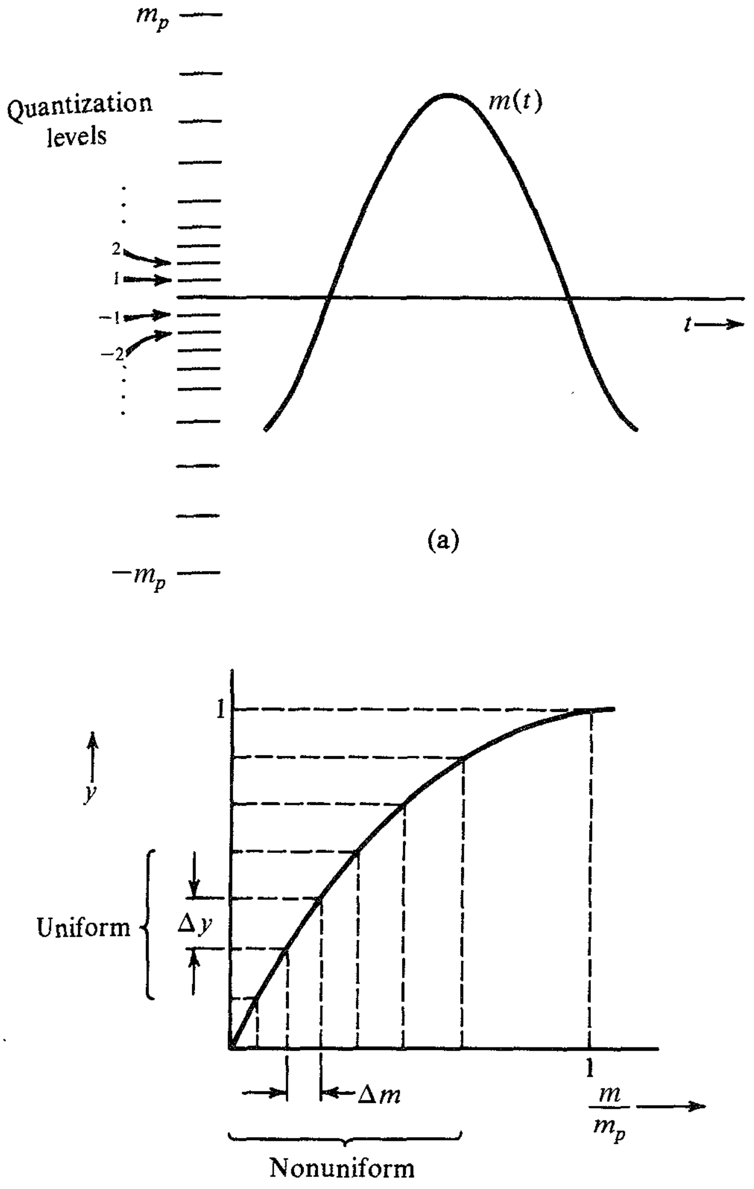
The output SNR for the cases of $\mu = 255$ and $\mu = 0$ (uniform quantization) as a function of $\widetilde{m^2}$ (the message signal power) is shown in Fig. 6.13

The Compandor

A logarithmic compressor can be realized by a semiconductor diode, because the V–I characteristic of such a diode is of the desired form in the first quadrant:

$$V = \frac{KT}{q} \ln\left(1 + \frac{I}{I_s}\right)$$

Two matched diodes in parallel with opposite polarity provide the approximate characteristic in the first and third quadrants (ignoring the saturation current). In practice, adjustable resistors are placed in series with each diode and a third variable resistor is added in parallel. By adjusting various resistors, the resulting characteristic is made to fit a finite number of points (usually seven) on the ideal characteristics.

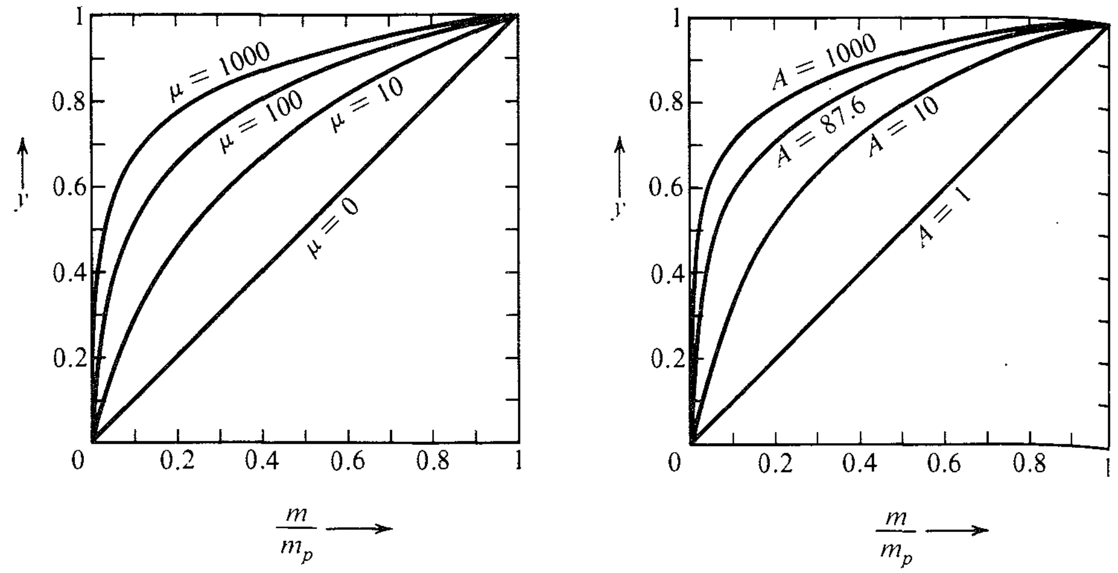


(a)

(b)

Figure 6.11 Nonuniform quantization.

An alternative approach is to use a piecewise linear approximation to the logarithmic characteristics. A 15-segmented approximation (Fig. 6.14) to the eight bit ($L = 256$) with $\mu = 255$ law is widely used in the D2 channel bank that is used in conjunction with the T1 carrier system. The segmented approximation is only marginally inferior in terms of SNR.[6] The piecewise linear approximation has almost universally replaced earlier logarithmic approximations to the true $\mu = 255$ characteristic and is the method of choice in North American standards. Though a true $\mu = 255$ compressor working with a $\mu = 255$ expander will be superior to similar piecewise linear devices, a digital terminal device exhibiting the true characteristic in today's network must work end-to-end against other network elements which use the piecewise linear approximation. Such a combination of differing characteristics



(a) (b)

Figure 6.12 (a) μ-law characteristic. (b) A-law characteristic.

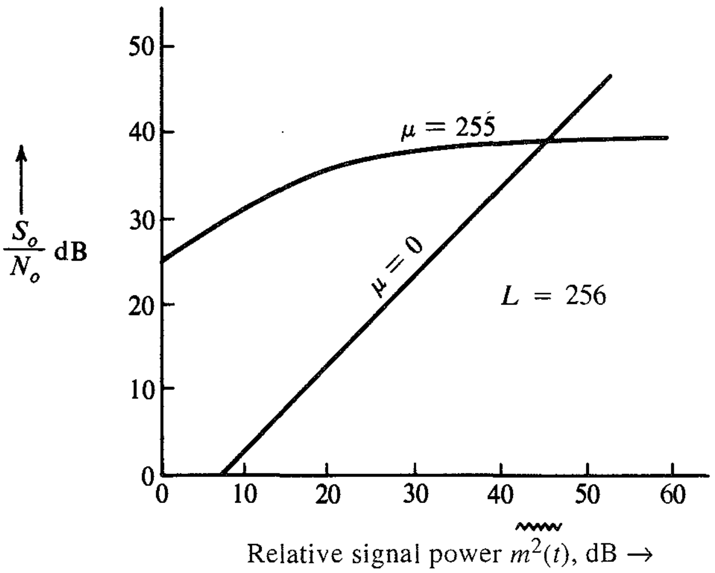


Figure 6.13 Signal-to-quantization-noise ratio in PCM with and without compression.

is inferior to either of the characteristics obtained when the compressor and the expander operate using the same compression law.

The Encoder

The multiplexed PAM output is applied at the input of the encoder, which quantizes and encodes each sample into a group of n binary digits. A variety of encoders is available.[7] We shall discuss here the **digit-at-a-time** encoder, which makes n sequential comparisons to generate an n-bit code word. The sample is compared with a voltage obtained by a combination of reference voltages proportional to $2^7, 2^6, 2^5, \ldots, 2^0$. The reference voltages are conveniently generated by a bank of resistors $R, 2R, 2^2R, \ldots, 2^7R$.

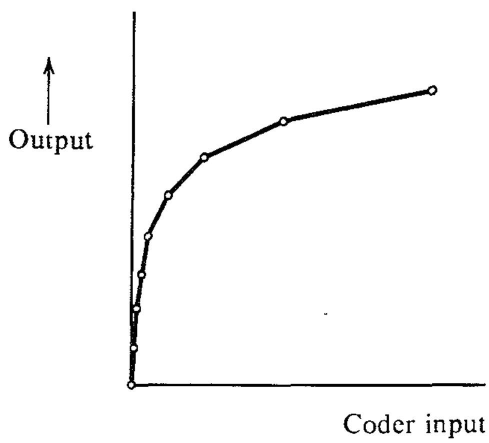


Figure 6.14 Piecewise linear compressor characteristic.

The encoding involves answering successive questions, beginning with whether or not the sample is in the upper or lower half of the allowed range. The first code digit **1** or **0** is generated, depending on whether the sample is in the upper or the lower half of the range. In the second step, another digit **1** or **0** is generated, depending on whether the sample is in the upper or the lower half of the subinterval in which it has been located. This process continues until the last binary digit in the code is generated.

Decoding is the inverse of encoding. In this case, each of the n digits is applied to a resistor of different value. The kth digit is applied to a resistor $2^k R$. The currents in all the resistors are added. The sum is proportional to the quantized sample value. For example, a binary code word **10010110** will give a current proportional to $2^7 + 0 + 0 + 2^4 + 0 + 2^2 + 2^1 + 0 = 150$. This completes the D/A conversion.

6.2.3 Transmission Bandwidth and the Output SNR

For a binary PCM, we assign a distinct group of n binary digits (bits) to each of the L quantization levels. Because a sequence of n binary digits can be arranged in 2^n distinct patterns,

$$L = 2^n \qquad \text{or} \qquad n = \log_2 L \tag{6.19}$$

Each quantized sample is, thus, encoded into n bits. Because a signal $m(t)$ band-limited to B Hz requires a minimum of $2B$ samples per second, we require a total of $2nB$ bits per second (bps), that is, $2nB$ pieces of information per second. Because a unit bandwidth (1 Hz) can transmit a maximum of two pieces of information per second (Sec. 6.1.3), we require a minimum channel of bandwidth B_T Hz, given by

$$B_T = nB \text{ Hz} \tag{6.20}$$

This is the theoretical minimum transmission bandwidth required to transmit the PCM signal. In Secs. 7.2 and 7.3, we shall see that for practical reasons we may use a transmission bandwidth higher than this.

EXAMPLE 6.2 A signal $m(t)$ band-limited to 3 kHz is sampled at a rate $33\frac{1}{3}\%$ higher than the Nyquist rate. The maximum acceptable error in the sample amplitude (the maximum quantization error) is

0.5% of the peak amplitude m_p. The quantized samples are binary coded. Find the minimum bandwidth of a channel required to transmit the encoded binary signal. If 24 such signals are time-division-multiplexed, determine the minimum transmission bandwidth required to transmit the multiplexed signal.

The Nyquist sampling rate is $R_N = 2 \times 3000 = 6000$ Hz (samples per second). The actual sampling rate is $R_A = 6000 \times (1\frac{1}{3}) = 8000$ Hz.

The quantization step is Δv, and the maximum quantization error is $\pm\Delta v/2$. Therefore, from Eq. (6.14),

$$\frac{\Delta v}{2} = \frac{m_p}{L} = \frac{0.5}{100} m_p \Longrightarrow L = 200$$

For binary coding, L must be a power of 2. Hence, the next higher value of L that is a power of 2 is $L = 256$.

From Eq. (6.19), we need $n = \log_2 256 = 8$ bits per sample. We require to transmit a total of $C = 8 \times 8000 = 64,000$ bit/s. Because we can transmit up to 2 bit/s per hertz of bandwidth, we require a minimum transmission bandwidth $B_T = C/2 = 32$ kHz.

The multiplexed signal has a total of $C_M = 24 \times 64,000 = 1.536$ Mbit/s, which requires a minimum of $1.536/2 = 0.768$ MHz of transmission bandwidth.

Exponential Increase of the Output SNR

From Eq. (6.19), $L^2 = 2^{2n}$, and the output SNR in Eq. (6.16) or Eq. (6.18) can be expressed as

$$\frac{S_o}{N_o} = c(2)^{2n} \tag{6.21}$$

where

$$c = \begin{cases} \dfrac{3\widetilde{m^2(t)}}{m_p^2} & \text{[uncompressed case, in Eq. (6.16)]} \\[2ex] \dfrac{3}{[\ln(1+\mu)]^2} & \text{[compressed case, in Eq. (6.18)]} \end{cases}$$

Substitution of Eq. (6.20) into Eq. (6.21) yields

$$\frac{S_o}{N_o} = c(2)^{2B_T/B} \tag{6.22}$$

From Eq. (6.22) we observe that the SNR increases exponentially with the transmission bandwidth B_T. This trade of SNR with bandwidth is attractive and comes close to the upper theoretical limit. A small increase in bandwidth yields a large benefit in terms of SNR. This relationship is clearly seen by rewriting Eq. (6.22) using the decibel scale as

$$\begin{aligned} \left(\frac{S_o}{N_o}\right)_{\text{dB}} &= 10 \log_{10}\left(\frac{S_o}{N_o}\right) \\ &= 10 \log_{10}[c(2)^{2n}] \\ &= 10 \log_{10} c + 2n \log_{10} 2 \\ &= (\alpha + 6n) \quad \text{dB} \end{aligned} \tag{6.23}$$

where $\alpha = 10 \log_{10} c$. This shows that increasing n by 1 (increasing one bit in the code word) quadruples the output SNR (6-dB increase). Thus, if we increase n from 8 to 9, the SNR quadruples, but the transmission bandwidth increases only from 32 to 36 kHz (an increase of only 12.5%). This shows that in PCM, SNR can be controlled by transmission bandwidth. We shall see later that frequency and phase modulation also do this. But it requires a doubling of the bandwidth to quadruple the SNR. In this respect, PCM is strikingly superior to FM or PM.

EXAMPLE 6.3 A signal $m(t)$ of bandwidth $B = 4$ kHz is transmitted using a binary companded PCM with $\mu = 100$. Compare the case of $L = 64$ with the case of $L = 256$ from the point of view of transmission bandwidth and the output SNR.

For $L = 64$, $n = 6$, and the transmission bandwidth is $nB = 24$ kHz,

$$\frac{S_o}{N_o} = (\alpha + 36) \text{ dB}$$

$$\alpha = 10 \log \frac{3}{[\ln (101)]^2} = -8.51$$

Hence,

$$\frac{S_o}{N_o} = 27.49 \text{ dB}$$

For $L = 256$, $n = 8$, and the transmission bandwidth is 32 kHz,

$$\frac{S_o}{N_o} = \alpha + 6n = 39.49 \text{ dB}$$

The difference between the two SNRs is 12 dB, which is a ratio of 16. Thus, the SNR for $L = 256$ is 16 times the SNR for $L = 64$. The former requires just about 33% more bandwidth compared to the latter.

Comments on Logarithmic Units

Logarithmic units and logarithmic scales are very convenient when a variable has a large dynamic range. Such is the case with frequency variables or SNRs. A logarithmic unit for the power ratio is the decibel (dB), defined as $10 \log_{10}$(power ratio). Thus, an SNR is x dB, where

$$x = 10 \log_{10} \frac{S}{N}$$

We use the same unit to express power gain or loss over a certain transmission medium. For instance, if over a certain cable the signal power is attenuated by a factor of 15, the cable gain is

$$G = 10 \log_{10} \frac{1}{15} = -11.76 \text{ dB}$$

or the cable attenuation (loss) is 11.76 dB.

Although the decibel is a measure of power ratios, it is often used as a measure of power itself. For instance, 100 watt power may be considered as a power ratio of 100 with respect to 1 watt power, and is expressed in units of dBW as

$$P_{\mathrm{dBW}} = 10 \log_{10} 100 = 20 \text{ dBW}$$

Thus, 100 watt power is 20 dBW. Similarly, power measured with respect to 1 mW power is dBm. For instance, 100 watt power is

$$P_{\mathrm{dBm}} = 10 \log \frac{100 \text{ W}}{1 \text{ mW}} = 50 \text{ dBm}$$

A Historical Note

A gap of more than 20 years occurred between the invention of PCM and its implementation because of the unavailability of suitable switching devices. Vacuum tubes, the devices available before the invention of the transistor, were not only bulky, but they were poor switches and dissipated a lot of power as heat. Systems using vacuum tubes as switches were large, rather unreliable, and tended to overheat. PCM was just waiting for the invention of the transistor, which happens to be small, consumes little power, and is a nearly ideal switch.

Coincidentally, at about the same time, the demand in telephone service had grown to the point where the old system was overloaded, particularly in large cities. It was not easy to install new underground cables because in many cities the space available under the streets was already occupied by other services (such as water, gas, sewer, etc.). Moreover, digging up streets and causing many dislocations was not very attractive. An attempt was made on a limited scale to increase the capacity by frequency-division-multiplexing several voice channels through amplitude modulation. Unfortunately, the cables were primarily designed for the audio voice range (0 to 4 kHz) and suffered severely from noise. Furthermore, cross talk at high frequencies between pairs of channels on the same cable was unacceptable. Ironically, PCM—requiring a bandwidth several times larger than that required for FDM signals—offered the solution. This is because digital systems with closely spaced regenerative repeaters can work satisfactorily on noisy, poor-high-frequency-performance lines. The repeaters, spaced approximately 6000 feet apart, clean up the signal and regenerate new pulses before the pulses get too distorted and noisy. This is the history of the Bell System's T1 carrier system.[3] A pair of wires that used to transmit one audio signal of bandwidth 4 kHz is now used to transmit 24 time-division-multiplexed PCM telephone signals with a total bandwidth of 1.544 MHz.

6.2.4 A T1 Carrier System

A schematic of a T1 carrier system is shown in Fig. 6.15a. All 24 channels are sampled in a sequence. The sampler output represents a time-division-multiplexed PAM signal. The multiplexed PAM signal is now applied to the input of an encoder that quantizes each sample and encodes it into eight binary pulses—a binary code word* (see Fig. 6.15b). The signal, now converted to digital form, is sent over the transmission medium. Regenerative repeaters spaced approximately 6000 feet apart detect the pulses and transmit new pulses. At the receiver, the decoder converts the binary pulses into samples (decoding). The samples are then demultiplexed (i.e., distributed to each of the 24 channels). The desired audio signal is reconstructed by passing the samples through a low-pass filter in each channel.

---

* In an earlier version, each sample was encoded by seven bits. An additional bit was added for signaling.

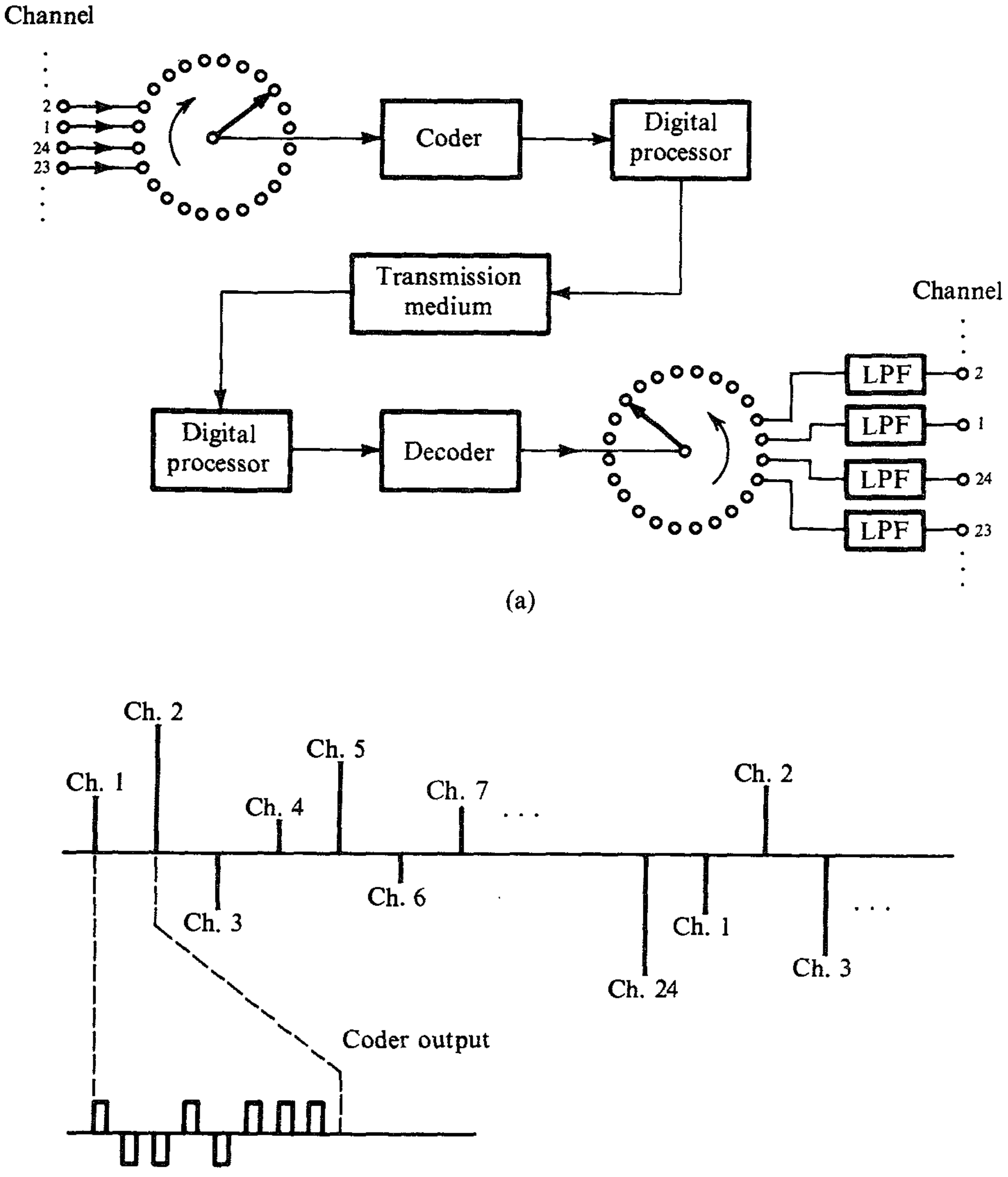

Channel
2
1
24
23
Coder
Digital processor
Transmission medium
Channel
Digital processor
Decoder
LPF
LPF
LPF
LPF
2
1
24
23
(a)
Ch. 2
Ch. 1
Ch. 5
Ch. 4
Ch. 7
Ch. 2
Ch. 3
Ch. 6
Ch. 1
Ch. 24
Ch. 3
Coder output

Figure 6.15 T1 carrier system.

The commutators in Fig. 6.15 are not mechanical but are high-speed electronic switching circuits. Several schemes are available for this purpose.[8] Sampling is done by electronic gates (such as a bridge diode circuit, as shown in Fig. 4.5) opened periodically by narrow pulses of 2-μs duration. The 1.544-Mbit/s signal of the T1 system is called **digital signal level 1 (DS1)**, which is used further to multiplex into progressively higher level signals DS2, DS3, and DS4, as described in Sec. 7.9.

After the Bell System introduced the T1 carrier system in the United States, dozens of variations were proposed or adopted elsewhere before the CCITT standardized its 30-channel PCM system with a rate of 2.048 Mbit/s (in contrast to T1, with 24 channels and 1.544 Mbit/s).

The 30-channel system is used all over the world, except in North America and Japan. Because of the widespread adoption of the T1 carrier system in the United States and Japan before the CCITT standardization, the two standards continue to be used in different parts of the world, with appropriate interfaces in international communication.

Synchronizing and Signaling

Binary code words corresponding to samples of each of the 24 channels are multiplexed in a sequence, as shown in Fig. 6.16. A segment containing one code word (corresponding to one sample) from each of the 24 channels is called a **frame**. Each frame has $24 \times 8 = 192$ information bits. Because the sampling rate is 8000 samples per second, each frame takes 125 μs. At the receiver, it is necessary to be sure where each frame begins in order to separate information bits correctly. For this purpose, a **framing bit** is added at the beginning of each frame. This makes a total of 193 bits per frame. Framing bits are chosen so that a sequence of framing bits, one at the beginning of each frame, forms a special pattern that is unlikely to be formed in a speech signal.

The sequence formed by the first bit from each frame is examined by the logic of the receiving terminal. If this sequence does not follow the given coded pattern (framing bit pattern), then a synchronization loss is detected, and the next position is examined to determine whether it is actually the framing bit. It takes about 0.4 to 6 ms to detect and about 50 ms (in the worst possible case) to reframe.

In addition to information and framing bits, we need to transmit signaling bits corresponding to dialing pulses, as well as telephone on-hook/off-hook signals. When channels

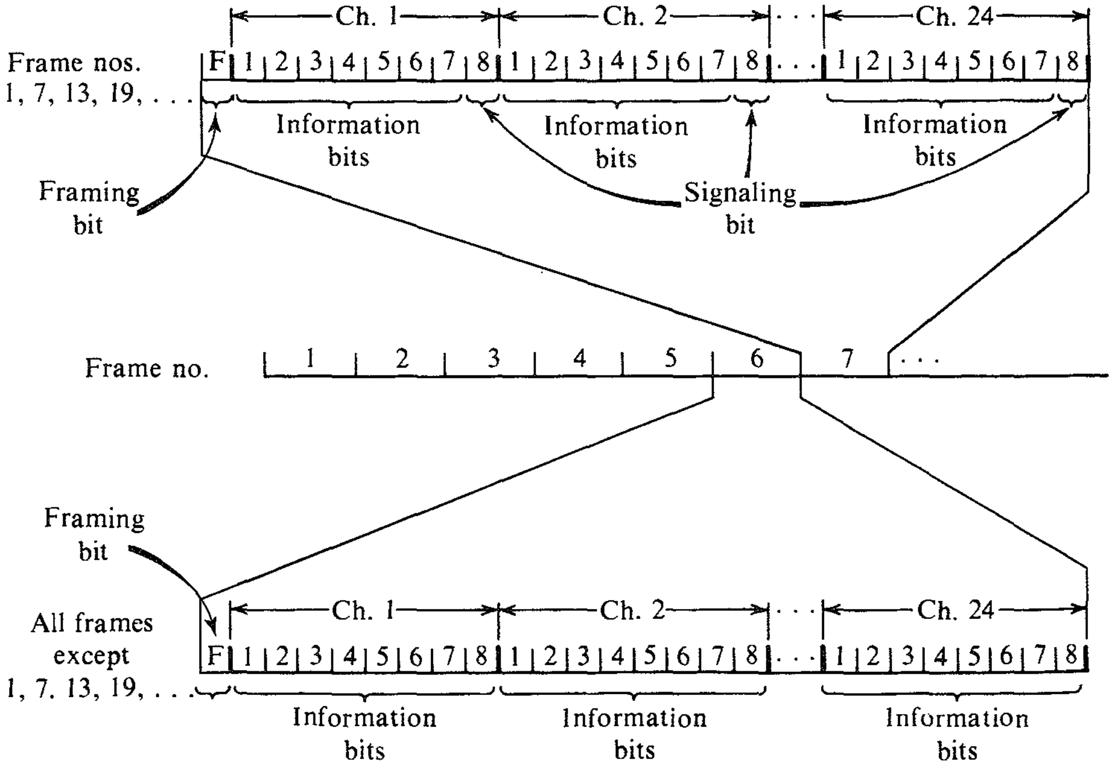

Ch. 1
Ch. 2
Ch. 24
Frame nos.
1, 7, 13, 19, . . .
F 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 . . . 1 2 3 4 5 6 7 8
Information bits
Information bits
Information bits
Framing bit
Signaling bit
Frame no.
1 2 3 4 5 6 7 . . .
Framing bit
All frames except 1, 7, 13, 19, . . .
F 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 . . . 1 2 3 4 5 6 7 8
Information bits
Information bits
Information bits

Figure 6.16 T1 system signaling format.

developed by this system are used to transmit signals between telephone switching systems, the switches must be able to communicate with each other to use the channels effectively. Since all eight bits are now used for transmission instead of the seven bits used in the earlier version,* the signaling channel provided by the eighth bit is no longer available. Since only a rather low-speed signaling channel is required, rather than create extra time slots for this information, we use one information bit (the least significant bit) of every sixth sample of a signal to transmit this information. This means every sixth sample of each voice signal will have a possible error corresponding to the least significant digit. Every sixth frame, therefore, has $7 \times 24 = 168$ information bits, 24 signaling bits, and 1 framing bit. In all the remaining frames, there are 192 information bits and 1 framing bit. This technique is called $\mathbf{7\frac{5}{6}}$ bit encoding, and the signaling channel so derived is called **robbed-bit signaling**. The slight SNR degradation suffered by impairing one out of six frames is considered to be an acceptable penalty. The signaling bits for each signal occur at a rate of $8000/6 = 1333$ bit/s. The frame format is shown in Fig. 6.16.

The older seven-bit framing format required only that frame boundaries be identified so that the channels could be located in the bit stream. When signaling is superimposed on the channels in every sixth frame, it is necessary to identify, at the receiver, which frames are the signaling frames. A new framing structure, called the **superframe**, was developed to take care of this. The framing bits are transmitted at the 8-kbit/s rate as before and occupy the first bit of each frame. The framing bits form a special pattern which repeats in twelve frames: **100011011100**. The pattern thus allows the identification of frame boundaries as before, but also allows the determination of the locations of the sixth and twelfth frames within the superframe. Note that the superframe described here is twelve frames in length. Since two bits per superframe are available for signaling for each channel, it is possible to provide four-state signaling for a channel by using the four possible patterns of the two signaling bits: **00, 01, 10,** and **11**. Although most switch-to-switch applications in the telephone network require only two-state signaling, three- and four-state signaling techniques are used in certain special applications.

Advances in digital electronics and in coding theory have made it unnecessary to use the full 8 kbit/s of the framing channel in a DS1 signal to perform the framing task. A new superframe structure, called the **extended superframe (ESF)** format, was introduced during the 1970s to take advantage of the reduced framing bandwidth requirement. An ESF is 24 frames in length and carries signaling bits in the eighth bit of each channel in frames 6, 12, 18, and 24. Sixteen-state signaling is thus possible and is sometimes used although, as with the superframe format, most applications require only two-state signaling.

The 8-kbit/s overhead (framing) capacity of the ESF signal is divided into three channels: 2 kbit/s for framing, 2 kbit/s for a cyclic redundancy check (CRC-6) error detection channel, and 4 kbit/s for a data channel. The highly reliable error checking provided by the CRC-6 pattern and the use of the data channel to transport information on signal performance as received by the distant terminal make ESF much more attractive to service providers than the older superframe format. More discussion of this topic can be found in chapter 8.

The 2 kbit/s framing channel of the ESF format carries the repetitive pattern **001011. . . .**, a pattern that repeats in 24 frames and is much less subject to counterfeiting than the patterns associated with the earlier formats.

---

* In the earlier version of T1, quantizing levels $L = 128$ required only seven information bits. The eighth bit was used for signaling.

For various reasons, including the development of the intelligent network switching nodes, the function of signaling is being transferred out from the channels that carry the messages or data signals to separate signaling networks called **common channel interoffice signaling (CCIS)** systems. The universal deployment of such systems over the next few years will significantly decrease the importance of robbed-bit signaling, and all eight bits of each message (or sample) will be transmitted in most applications.

The Conference on European Postal and Telegraph Administration (CEPT) has standardized a PCM with 256 time slots per frame. Each frame has $30 \times 8 = 240$ information bits, corresponding to 30 speech channels (with eight bits each). The remaining 16 bits per frame are used for frame synchronization and signaling. Therefore, although the bit rate is 2.048 Mbit/s, corresponding to 32 voice channels, only 30 voice channels are transmitted.

6.3 DIFFERENTIAL PULSE-CODE MODULATION (DPCM)

In analog messages we can make a good guess about a sample value from a knowledge of the past sample values. In other words, the sample values are not independent, and generally there is a great deal of redundancy in the Nyquist samples. Proper exploitation of this redundancy leads to encoding a signal with a lesser number of bits. Consider a simple scheme where instead of transmitting the sample values, we transmit the difference between the successive sample values. Thus, if $m[k]$ is the kth sample, instead of transmitting $m[k]$, we transmit the difference $d[k] = m[k] - m[k-1]$. At the receiver, knowing $d[k]$ and the previous sample value $m[k-1]$, we can reconstruct $m[k]$. Thus, from the knowledge of the difference $d[k]$, we can reconstruct $m[k]$ iteratively at the receiver. Now, the difference between successive samples is generally much smaller than the sample values. Thus, the peak amplitude m_p of the transmitted values is reduced considerably. Because the quantization interval $\Delta v = m_p/L$, for a given L (or n), this reduces the quantization interval Δv, thus reducing the quantization noise, which is given by $\Delta v^2/12$. This means that for a given n (or transmission bandwidth), we can increase the SNR, or for a given SNR, we can reduce n (or transmission bandwidth).

We can improve upon this scheme by estimating (predicting) the value of the kth sample $m[k]$ from a knowledge of the previous sample values. If this estimate is $\hat{m}[k]$, then we transmit the difference (prediction error) $d[k] = m[k] - \hat{m}[k]$. At the receiver also, we determine the estimate $\hat{m}[k]$ from the previous sample values, and then generate $m[k]$ by adding the received $d[k]$ to the estimate $\hat{m}[k]$. Thus, we reconstruct the samples at the receiver iteratively. If our prediction is worth its salt, the predicted (estimated) value $\hat{m}[k]$ will be close to $m[k]$, and their difference (prediction error) $d[k]$ will be even smaller than the difference between the successive samples. Consequently, this scheme, known as the **differential PCM (DPCM)**, is superior to that described in the previous paragraph, which is a special case of DPCM, where the estimate of a sample value is taken as the previous sample value, that is, $\hat{m}[k] = m[k-1]$.

Spirits of Taylor, Maclaurin, and Wiener

Before describing DPCM, we shall briefly discuss the approach to signal prediction (estimation). To an uninitiated, future prediction seems a mysterious stuff fit only for psychics, wizards, mediums, and the likes, who can summon help from the spirit world. Electrical engineers appear to be hopelessly outclassed in this pursuit. Not quite so! We can also summon the spirits of Taylor, Maclaurin, Wiener, and the likes to help us. Consider, for example, a signal $m(t)$, which has derivatives of all orders at t. Using the Taylor series for this signal, we can express $m(t + T_s)$ as

$$m(t + T_s) = m(t) + T_s\dot{m}(t) + \frac{T_s^{\,2}}{2!}\ddot{m}(t) + \frac{T_s^{\,3}}{3!}\dddot{m}(t) + \cdots \tag{6.24a}$$

$$\approx m(t) + T_s\dot{m}(t) \qquad \text{for small } T_s \tag{6.24b}$$

Equation (6.24a) shows that from a knowledge of the signal and its derivatives at instant t, we can predict a future signal value at $t + T_s$. In fact, even if we know just the first derivative, we can still predict this value approximately, as shown in Eq. (6.24b). Let us denote the kth sample of $m(t)$ by $m[k]$, that is, $m(kT_s) = m[k]$, and $m(kT_s \pm T_s) = m[k \pm 1]$, and so on. Setting $t = kT_s$ in Eq. (6.24b), and recognizing that $\dot{m}(kT_s) \approx [m(kT_s) - m(kT_s - T_s)]/T_s$, we obtain

$$\begin{aligned} m[k+1] &\approx m[k] + T_s\left[\frac{m[k] - m[k-1]}{T_s}\right] \\ &= 2m[k] - m[k-1] \end{aligned}$$

This shows that we can find a crude prediction of the $(k + 1)$th sample from the two previous samples. The approximation in Eq. (6.24b) improves as we add more terms in the series on the right-hand side. To determine the higher order derivatives in the series, we require more samples in the past. The larger the number of past samples we use, the better will be the prediction. Thus, in general, we can express the prediction formula as

$$m[k] \approx a_1m[k-1] + a_2m[k-2] + \cdots + a_Nm[k-N] \tag{6.25}$$

The right-hand side is $\hat{m}[k]$, the predicted value of $m[k]$. Thus,

$$\hat{m}[k] = a_1m[k-1] + a_2m[k-2] + \cdots + a_Nm[k-N] \tag{6.26}$$

This is the equation of an Nth-order predictor. Larger N would result in better prediction in general. The output of this filter (predictor) is $\hat{m}[k]$, the predicted value of $m[k]$. The input is the previous samples $m[k-1]$, $m[k-2]$, $\cdots$, $m[k-N]$, although it is customary to say that the input is $m[k]$ and the output is $\hat{m}[k]$. Observe that this equation reduces to $\hat{m}[k] = m[k-1]$ for the first-order predictor. It follows from Eq. (6.24b), where we retain only the first term on the right-hand side. This means that $a_1 = 1$, and the first-order predictor is a simple time delay.

We have outlined here a very simple procedure for predictor design. In a more sophisticated approach, discussed in Sec. 10.6, where we use the minimum mean squared error criterion for best prediction, the **prediction coefficients** a_j in Eq. (6.26) are determined from the statistical correlation between various samples. The predictor described in Eq. (6.26) is called a **linear predictor**. It is basically a transversal filter (a tapped delay line), where the tap gains are set equal to the prediction coefficients, as shown in Fig. 6.17.

Analysis of DPCM

As mentioned earlier, in DPCM we transmit not the present sample $m[k]$, but $d[k]$ (the difference between $m[k]$ and its predicted value $\hat{m}[k]$). At the receiver, we generate $\hat{m}[k]$ from the past sample values to which the received $d[k]$ is added to generate $m[k]$. There is, however, one difficulty in this scheme. At the receiver, instead of the past samples $m[k-1]$, $m[k-2]$, $\cdots$

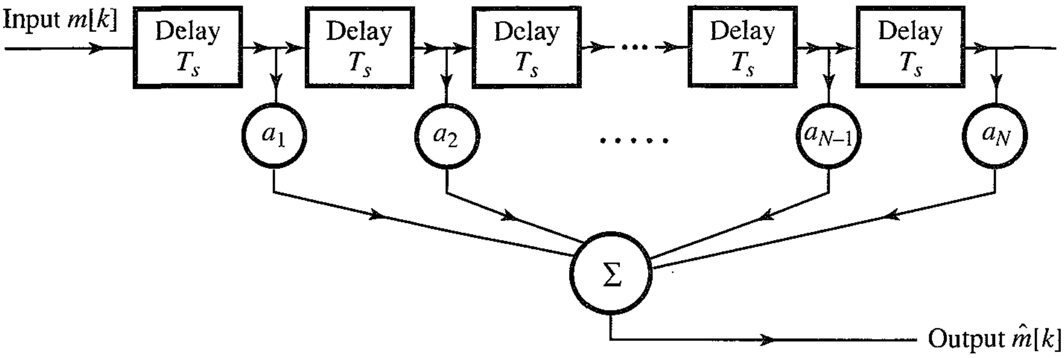


Figure 6.17 Transversal filter (tapped delay line) used as a linear predictor.

as well as $d[k]$, we have their quantized versions $m_q[k-1]$, $m_q[k-2]$, $\cdots$. Hence, we cannot determine $\hat{m}[k]$. We can only determine $\hat{m}_q[k]$, the estimate of the quantized sample $m_q[k]$, in terms of the quantized samples $m_q[k-1]$, $m_q[k-2]$, $\cdots$. This will increase the error in reconstruction. In such a case, a better strategy is to determine $\hat{m}_q[k]$, the estimate of $m_q[k]$ (instead of $m[k]$), at the transmitter also from the quantized samples $m_q[k-1]$, $m_q[k-2]$, $\cdots$. The difference $d[k] = m[k] - \hat{m}_q[k]$ is now transmitted using PCM. At the receiver, we can generate $\hat{m}_q[k]$, and from the received $d[k]$, we can reconstruct $m_q[k]$.

Figure 6.18a shows a DPCM transmitter. We shall soon show that the predictor input is $m_q[k]$. Naturally, its output is $\hat{m}_q[k]$, the predicted value of $m_q[k]$. The difference

$$d[k] = m[k] - \hat{m}_q[k] \tag{6.27}$$

is quantized to yield

$$d_q[k] = d[k] + q[k] \tag{6.28}$$

where $q[k]$ is the quantization error. The predictor output $\hat{m}_q[k]$ is fed back to its input so that the predictor input $m_q[k]$ is

$$\begin{aligned} m_q[k] &= \hat{m}_q[k] + d_q[k] \\ &= m[k] - d[k] + d_q[k] \\ &= m[k] + q[k] \end{aligned} \tag{6.29}$$

This shows that $m_q[k]$ is a quantized version of $m[k]$. The predictor input is indeed $m_q[k]$, as assumed. The quantized signal $d_q[k]$ is now transmitted over the channel. The receiver shown in Fig. 6.18b is identical to the shaded portion of the transmitter. The inputs in both cases are also the same, viz., $d_q[k]$. Therefore, the predictor output must be $\hat{m}_q[k]$ (the same as the predictor output at the transmitter). Hence, the receiver output (which is the predictor input) is also the same, viz., $m_q[k] = m[k] + q[k]$, as found in Eq. (6.29). This shows that we are able to receive the desired signal $m[k]$ plus the quantization noise $q[k]$. This is the quantization noise associated with the difference signal $d[k]$, which is generally much smaller than $m[k]$. The received samples are decoded and passed through a low-pass filter for D/A conversion.

SNR Improvement

To determine the improvement in DPCM over PCM, let m_p and d_p be the peak amplitudes of $m(t)$ and $d(t)$, respectively. If we use the same value of L in both cases, the quantization

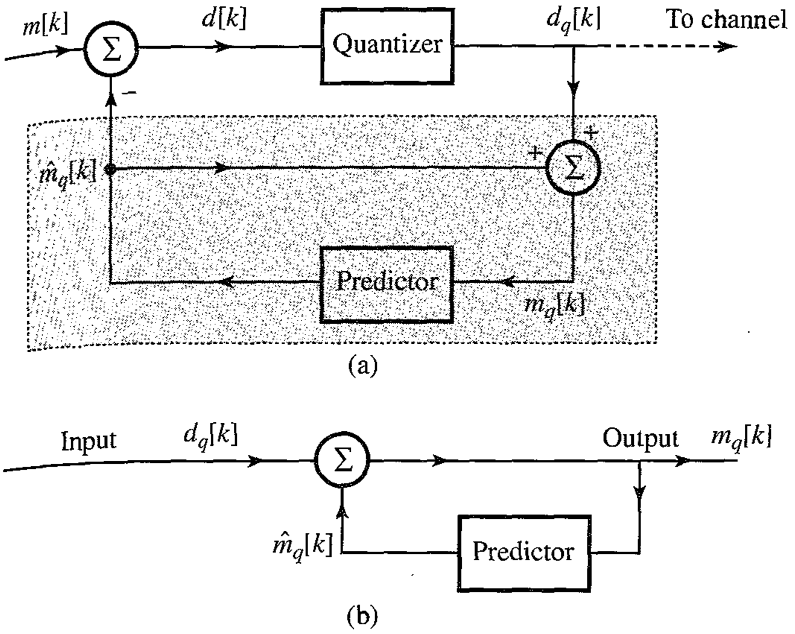


Figure 6.18 DPCM system. (a) Transmitter. (b) Receiver.

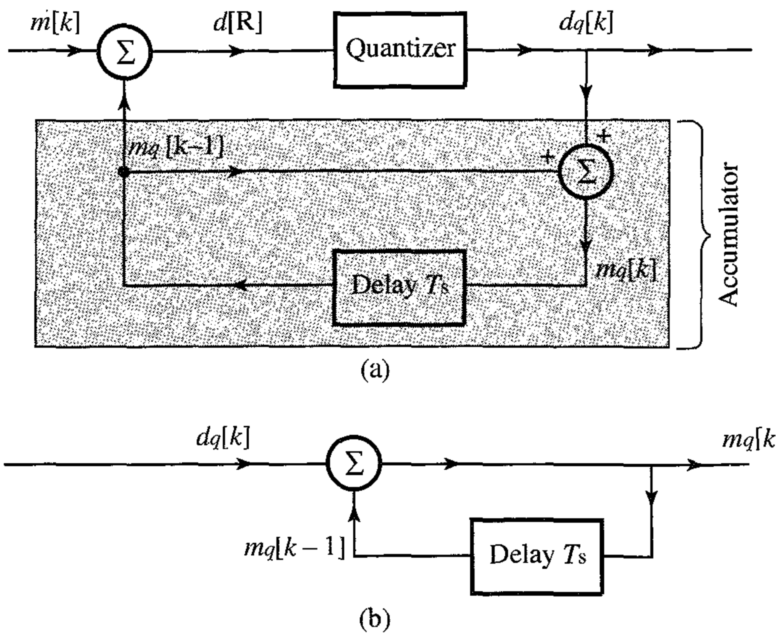
step Δv in DPCM is reduced by the factor d_p/m_p. Because the quantization noise power is $(\Delta v)^2/12$, the quantization noise in DPCM reduces by the factor $(m_p/d_p)^2$, and the SNR increases by the same factor. Moreover, the signal power is proportional to its peak value squared (assuming other statistical properties invariant). Therefore, G_p (SNR improvement due to prediction) is

$$G_p = \frac{P_m}{P_d}$$

where P_m and P_d are the powers of $m(t)$ and $d(t)$, respectively. In terms of dB units, this means that the SNR increases by $10 \log_{10}(P_m/P_d)$ dB. Therefore, Eq. (6.23) applies to DPCM also with a value of α that is higher by $10 \log_{10}(P_m/P_d)$ dB. In Example 10.21, a second-order predictor processor for speech signals is analyzed. For this case, the SNR improvement is found to be 5.6 dB. In practice, the SNR improvement may be as high as 25 dB in such cases as short-term voiced speech spectra and in the spectra of low-activity images.[9] Alternately, for the same SNR, the bit rate for DPCM could be lower than that for PCM by 3 to 4 bits per sample. Thus, telephone systems using DPCM can often operate at 32 kbit/s or even 24 kbit/s.

6.4 DELTA MODULATION

Sample correlation used in DPCM is further exploited in **delta modulation (DM)** by oversampling (typically 4 times the Nyquist rate) the baseband signal. This increases the correlation between adjacent samples, which results in a small prediction error that can be encoded using only one bit ($L = 2$). Thus, DM is basically a 1-bit DPCM, that is, a DPCM that uses only two levels ($L = 2$) for quantization of the $m[k] - \hat{m}_q[k]$. In comparison to PCM (and DPCM), it is a very simple and inexpensive method of A/D conversion. A 1-bit code word in DM makes word framing unnecessary at the transmitter and the receiver. This strategy allows us to use fewer bits per sample for encoding a baseband signal.



(a)

(b)

Figure 6.19 Delta modulation is a special case of DPCM.

In DM, we use a first-order predictor, which, as seen earlier, is just a time delay of T_s (the sampling interval). Thus, the DM transmitter (modulator) and receiver (demodulator) are identical to those of the DPCM in Fig. 6.18, with a time delay for the predictor, as shown in Fig. 6.19. From this figure, we obtain

$$m_q[k] = m_q[k-1] + d_q[k] \tag{6.30}$$

Hence,

$$m_q[k-1] = m_q[k-2] + d_q[k-1]$$

Substituting this equation into Eq. (6.30) yields

$$m_q[k] = m_q[k-2] + d_q[k] + d_q[k-1]$$

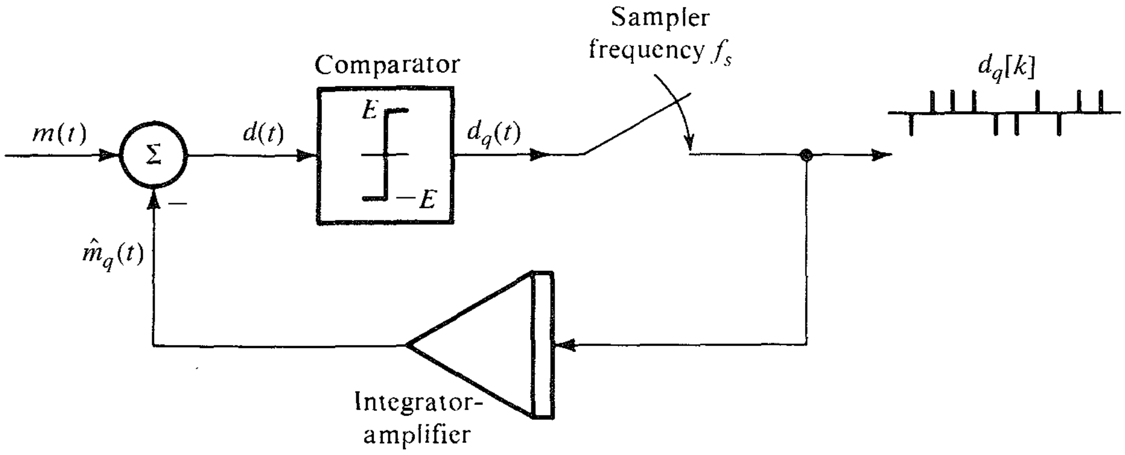
Proceeding iteratively in this manner, and assuming zero initial condition, that is, $m_q[0] = 0$, yields

$$m_q[k] = \sum_{m=0}^{k} d_q[m] \tag{6.31}$$

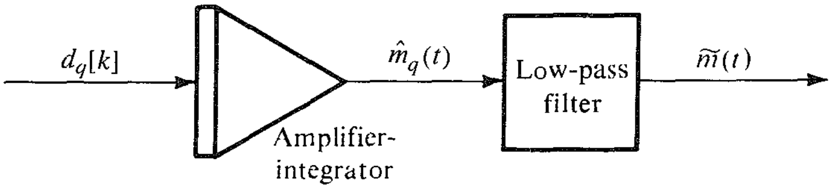
This shows that the receiver (demodulator) is just an accumulator (adder). If the output $d_q[k]$ is represented by impulses, then the accumulator (receiver) may be realized by an integrator because its output is the sum of the strengths of the input impulses (sum of the areas under the impulses). We may also replace the feedback portion of the modulator (which is identical to the demodulator) by an integrator. The demodulator output is $m_q[k]$, which when passed through a low-pass filter yields the desired signal reconstructed from the quantized samples.

Figure 6.20 shows a practical implementation of the delta modulator and demodulator. As discussed earlier, the first-order predictor is replaced by a low-cost integrator circuit (such as an RC integrator). The modulator (Fig. 6.20a) consists of a comparator and a sampler in the direct path and an integrator-amplifier in the feedback path. Let us see how this delta modulator works.

The analog signal $m(t)$ is compared with the feedback signal (which serves as a predicted signal) $\hat{m}_q(t)$. The error signal $d(t) = m(t) - \hat{m}_q(t)$ is applied to a comparator. If $d(t)$ is


Sampler frequency f_s
Comparator
$d_q[k]$
$m(t)$
Σ
$d(t)$
E
$-E$
$d_q(t)$
−
$\hat{m}_q(t)$
Integrator-amplifier

(a) Delta modulator


$d_q[k]$
$\hat{m}_q(t)$
Low-pass filter
$\tilde{m}(t)$
Amplifier-integrator

(b) Delta demodulator

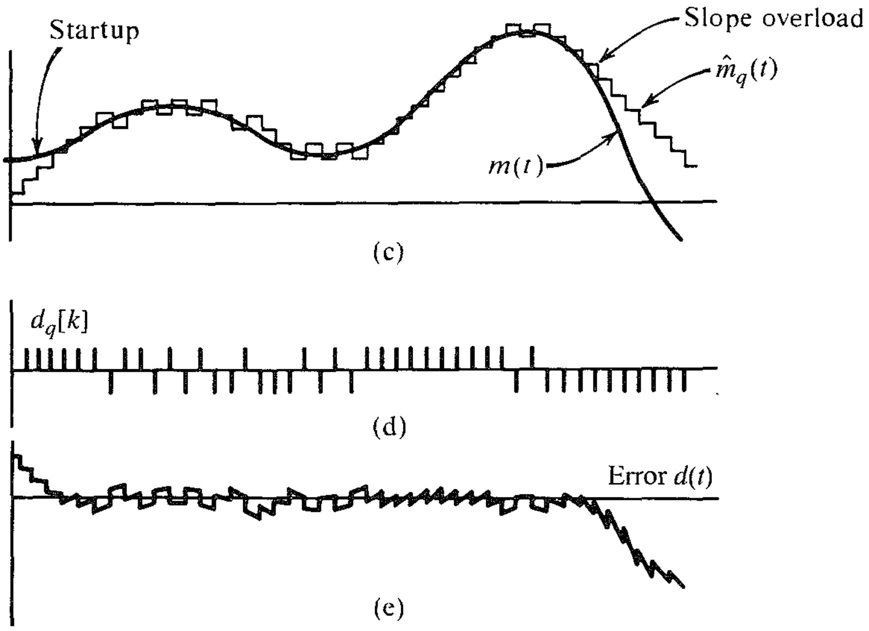


Figure 6.20 Delta modulation.

positive, the comparator output is a constant signal of amplitude E, and if $d(t)$ is negative, the comparator output is $-E$. Thus, the difference is a binary signal ($L = 2$) that is needed to generate a 1-bit DPCM. The comparator output is sampled by a sampler at a rate of f_s samples per second, where f_s is typically much higher than the Nyquist rate. The sampler thus produces a train of narrow pulses $d_q[k]$ (to simulate impulses) with a positive pulse when $m(t) > \hat{m}_q(t)$ and a negative pulse when $m(t) < \hat{m}_q(t)$. Note that each sample is coded by a single binary pulse (1-bit DPCM), as required. The pulse train $d_q[k]$ is the delta-modulated pulse train (Fig. 6.20d). The modulated signal $d_q[k]$ is amplified and integrated in the feedback path to generate $\hat{m}_q(t)$ (Fig. 6.20c), which tries to follow $m(t)$.

To understand how this works we note that each pulse in $d_q[k]$ at the input of the integrator gives rise to a step function (positive or negative, depending on the pulse polarity) in $\hat{m}_q(t)$. If, for example, $m(t) > \hat{m}_q(t)$, a positive pulse is generated in $d_q[k]$, which gives rise to a positive step in $\hat{m}_q(t)$, trying to equalize $\hat{m}_q(t)$ to $m(t)$ in small steps at every sampling instant, as shown in Fig. 6.20c. It can be seen that $\hat{m}_q(t)$ is a kind of staircase approximation of $m(t)$. When $\hat{m}_q(t)$ is passed through a low-pass filter, the coarseness of the staircase in $\hat{m}_q(t)$ is eliminated, and we get a smoother and better approximation to $m(t)$. The demodulator at the receiver consists of an amplifier-integrator (identical to that in the feedback path of the modulator) followed by a low-pass filter (Fig. 6.20b).

DM Transmits the Derivative of $m(t)$

In PCM, the analog signal samples are quantized in L levels, and this information is transmitted by n pulses per sample ($n = \log_2 L$). A little reflection shows that in DM, the modulated signal carries information not about the signal samples but about the difference between successive samples. If the difference is positive or negative, a positive or a negative pulse (respectively) is generated in the modulated signal $d_q[k]$. Basically, therefore, DM carries the information about the derivative of $m(t)$ and, hence, the name delta modulation. This can also be seen from the fact that integration of the delta-modulated signal yields $\hat{m}_q(t)$, which is an approximation of $m(t)$.

In PCM, the information of each quantized sample is transmitted by an n-bit code word, whereas in DM the information of the difference between successive samples is transmitted by a 1-bit code word.

Threshold of Coding and Overloading

Threshold and overloading effects can be clearly seen in Fig. 6.20c. Variations in $m(t)$ smaller than the step value (threshold of coding) are lost in DM. Moreover, if $m(t)$ changes too fast, that is, $\dot{m}(t)$ is too high, $\hat{m}_q(t)$ cannot follow $m(t)$, and overloading occurs. This is the so-called **slope overload**, which gives rise to the slope overload noise. This noise is one of the basic limiting factors in the performance of DM. We should expect slope overload rather than amplitude overload in DM, because DM basically carries the information about $\dot{m}(t)$. The granular nature of the output signal gives rise to the granular noise similar to the quantization noise. The slope overload noise can be reduced by increasing σ (the step size). This unfortunately increases the granular noise. There is an optimum value of σ, which yields the best compromise giving the minimum overall noise. This optimum value of σ depends on the sampling frequency f_s and the nature of the signal.[9]

The slope overload occurs when $\hat{m}_q(t)$ cannot follow $m(t)$. During the sampling interval T_s, $\hat{m}_q(t)$ is capable of changing by σ, where σ is the height of the step. Hence, the maximum slope that $\hat{m}_q(t)$ can follow is σ/T_s, or σf_s, where f_s is the sampling frequency. Hence, no overload occurs if

$$|\dot{m}(t)| < \sigma f_s$$

Consider the case of tone modulation (meaning a sinusoidal message):

$$m(t) = A\cos\omega t$$

The condition for no overload is

$$|\dot{m}(t)|_{\text{max}} = \omega A < \sigma f_s \tag{6.32}$$

Hence, the maximum amplitude A_{max} of this signal that can be tolerated without overload is given by

$$A_{\text{max}} = \frac{\sigma f_s}{\omega} \tag{6.33}$$

The overload amplitude of the modulating signal is inversely proportional to the frequency ω. For higher modulating frequencies, the overload occurs for smaller amplitudes. For voice signals, which contain all frequency components up to (say) 4 kHz, calculating A_{max} by using $\omega = 2\pi \times 4000$ in Eq. (6.33) will give an overly conservative value. It has been shown by de Jager[10] that A_{max} for voice signals can be calculated by using $\omega_r \simeq 2\pi \times 800$ in Eq. (6.33),

$$[A_{\text{max}}]_{\text{voice}} \simeq \frac{\sigma f_s}{\omega_r} \tag{6.34}$$

Thus, the maximum voice signal amplitude A_{max} that can be used without causing slope overload in DM is the same as the maximum amplitude of a sinusoidal signal of reference frequency f_r ($f_r \simeq 800$ Hz) that can be used without causing slope overload in the same system.

Fortunately, the voice spectrum (as well as the television video signal) also decays with frequency and closely follows the overload characteristics (curve *c*, Fig. 6.21). For this reason, DM is well suited for voice (and television) signals. Actually, the voice signal spectrum (curve *b*) decreases as $1/\omega$ up to 2000 Hz, and beyond this frequency, it decreases as $1/\omega^2$. If we had used a double integration in the feedback circuit instead of a single integration, A_{max} in Eq. (6.33) would be proportional to $1/\omega^2$. Hence, a better match between the voice spectrum and the overload characteristics is achieved by using a single integration up to 2000 Hz and a double integration beyond 2000 Hz. Such a circuit (the double integration) is fast responding, but has a tendency to instability, which can be reduced by using some low-order prediction along with double integration. The double integrator can be built by placing in cascade two low-pass *RC* integrators with the time constants $R_1C_1 = 1/200\pi$ and $R_2C_2 = 1/4000\pi$, respectively. This results in single integration from 100 Hz to 2000 Hz and double integration beyond 2000 Hz.

Adaptive Delta Modulation (ADM)

The DM discussed so far suffers from one serious disadvantage. The dynamic range of amplitudes is too small because of the threshold and overload effects discussed earlier. To correct this problem, some type of signal compression is necessary. In DM, a suitable method appears to be the adaptation of the step value σ according to the level of the input signal

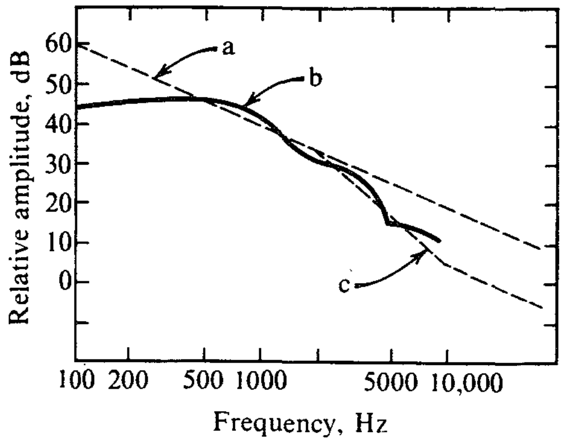


Figure 6.21 Voice signal spectrum.

derivative. For example, in Fig. 6.20, when the signal $m(t)$ is falling rapidly, slope overload occurs. If we can increase the step size during this period, the overload could be avoided. On the other hand, if the slope of $m(t)$ is small, a reduction of step size will reduce the threshold level as well as the granular noise. The slope overload causes $d_q[k]$ to have several pulses of the same polarity in succession. This calls for increased step size. Similarly, pulses in $d_q[k]$ alternating continuously in polarity indicates small amplitude variations, requiring a reduction in step size. In ADM we detect such pulse patterns and automatically adjust the step size.[11] This results in a much larger dynamic range for DM.

Output SNR

The error $d(t)$ caused by the granular noise in DM, shown in Fig. 6.20e (excluding slope overload portion), lies in the range $(-\sigma, \sigma)$, where σ is the step height in $\hat{m}_q(t)$. The situation is similar to that encountered in PCM, where the quantization-error amplitude was in the range from $-\Delta v/2$ to $\Delta v/2$. Hence, from Eq. (6.15a), the granular noise power $\overline{\epsilon^2}$ is

$$\overline{\epsilon^2} = \frac{\sigma^2}{3} \tag{6.35}$$

The granular noise PSD has a continuous spectrum, with most of the power in the frequency range extending well beyond the sampling frequency f_s. At the output, most of this will be suppressed by the baseband filter of bandwidth B. Hence, the granular noise power N_o will be well below that indicated in Eq. (6.35). To compute N_o, we shall make the assumption that the PSD of the quantization noise is uniform and concentrated in the band of 0 to f_s Hz. This assumption has been verified experimentally. Because the total power $\sigma^2/3$ is uniformly spread over the bandwidth f_s, the power within the baseband B is

$$N_o = \left(\frac{\sigma^2}{3}\right)\frac{B}{f_s} = \frac{\sigma^2 B}{3f_s} \tag{6.36}$$

The output signal power is $S_o = \widetilde{m^2(t)}$. Assuming no slope overload distortion,

$$\frac{S_o}{N_o} = \frac{3f_s\widetilde{m^2(t)}}{\sigma^2 B} \tag{6.37}$$

If m_p is the peak signal amplitude, then from Eq. (6.34),

$$m_p = \frac{\sigma f_s}{\omega_r} \tag{6.38}$$

and

$$\frac{S_o}{N_o} = \frac{3 f_s^3 \widetilde{m^2(t)}}{\omega_r^2 B m_p^2} \tag{6.39}$$

Because we need to transmit f_s pulses per second, the minimum transmission bandwidth $B_T = f_s/2$. Also for voice signals, $B = 4000$ and $\omega_r = 2\pi \times 800 = 1600\pi$. Hence,

$$\frac{S_o}{N_o} = \frac{150}{\pi^2}\left(\frac{B_T}{B}\right)^3 \frac{\widetilde{m^2(t)}}{m_p^2} \tag{6.40}$$

Thus, the output SNR varies as the cube of the bandwidth expansion ratio B_T/B. This result is derived for the single-integration case. For double-integration DM, Greefkes and de Jager have shown that[12]

$$\frac{S_o}{N_o} = 5.34\left(\frac{B_T}{B}\right)^5 \frac{\widetilde{m^2(t)}}{m_p^2} \tag{6.41}$$

It should be remembered that these results are valid only for voice signals. In all the preceding developments, we have ignored the pulse detection error at the receiver.

Comparison with PCM

The SNR in DM varies as a power of B_T/B, being proportional to $(B_T/B)^3$ for single integration and $(B_T/B)^5$ for double integration. In PCM, on the other hand, the SNR varies exponentially with B_T/B. Whatever the initial value, the exponential will always outrun the power variation. Clearly for higher values of B_T/B, PCM is expected to be superior to DM. The output SNR for voice signals as a function of the bandwidth expansion ratio B_T/B is plotted in Fig. 6.22 for tone modulation for which $\widetilde{m^2}/m_p^2 = 0.5$. The transmission bandwidth is assumed to be the theoretical minimum bandwidth for DM as well as PCM. From this figure it is clear that DM with double integration has a performance superior to companded PCM (which is the practical case) for lower values of B_T/B, and PCM is superior to DM for higher B_T/B, the crossover occurring at about $B_T/B = 10$. In practice, the crossover value is lower than 10, usually between 6 and 7 ($f_s \approx 50$ kbit/s).[13] This is true only of voice and television signals, for which DM is ideally suited. For other types of signals, DM does not compare as well with PCM.

Because the DM signal is a digital signal, it has all the advantages of digital systems, such as the use of regenerative repeaters and other advantages mentioned earlier. As far as detection errors are concerned, DM is more immune to this kind of error than is PCM, where the weight of the detection error depends on the digit location; thus for $n = 8$, the error in the first digit is 128 times as large as the error in the last digit. For DM, on the other hand, each digit has equal importance. Experiments have shown that an error probability P_e on the order of 10^{-1} does not affect the intelligibility of voice signals in DM, whereas P_e as low as 10^{-4} can cause serious error, leading to threshold in PCM. For multiplexing several channels, however, DM suffers from the fact that each channel requires its own coder and decoder, whereas for PCM, one coder and one decoder are shared by all channels. But this very fact of an individual coder and

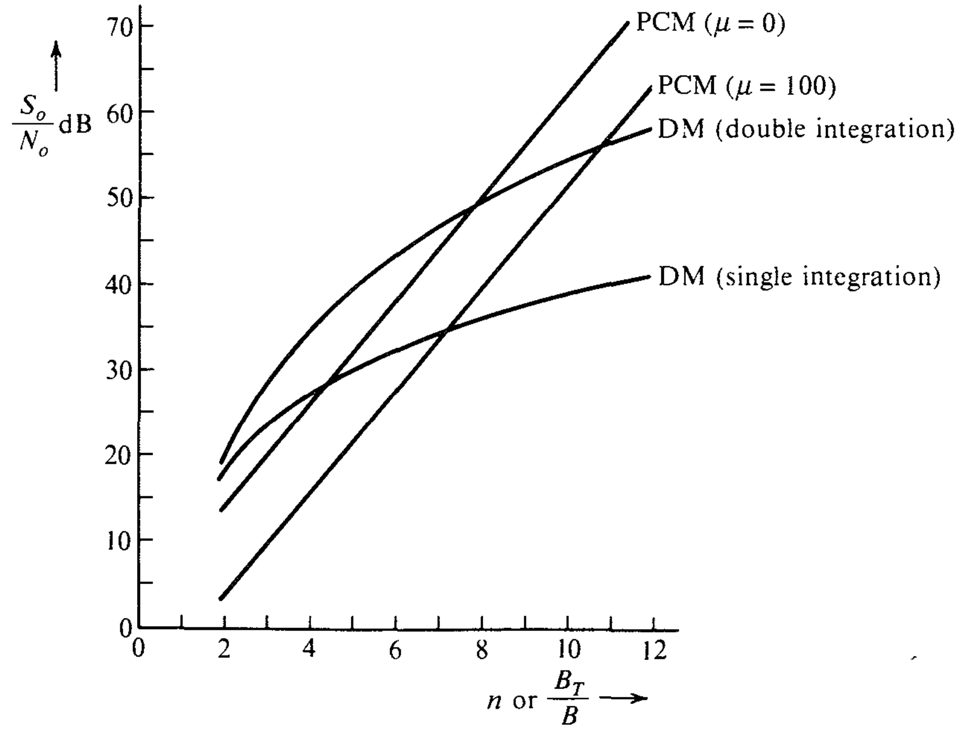


Figure 6.22 Comparison of DM and PCM.

decoder for each channel also permits more flexibility in DM. On the route between terminals, it is easy to drop one or more channels and insert other incoming channels. For PCM, such operations can be performed only at the terminals. This is particularly attractive for rural areas with low population density and where the population grows progressively. The individual coder-decoder also avoids cross talk, thus alleviating the stringent design requirements in the multiplexing circuits in PCM.

In conclusion, DM can outperform PCM at low SNR, but is inferior to PCM in the high SNR case. One of the advantages of DM is its simplicity, which also makes it less expensive. However, the cost of digital components, including A/D converters, is coming down to the point that the cost advantage of DM becomes insignificant.

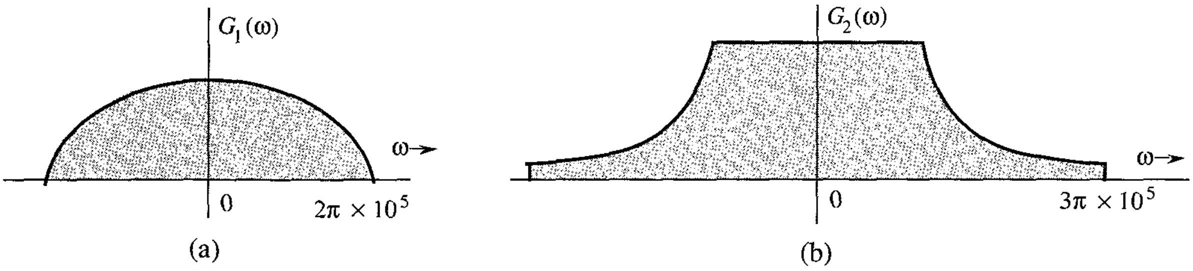
REFERENCES

1. D. A. Linden, "A Discussion of Sampling Theorem," *Proc. IRE*, vol. 47, pp. 1219-1226, July 1959.
2. R. N. Bracewell, *The Fourier Transform and Its Applications*, 2nd rev. ed., McGraw-Hill, New York, 1986.
3. W. R. Bennett, *Introduction to Signal Transmission*, McGraw-Hill, New York, 1970.
4. B. Smith, "Instantaneous Companding of Quantized Signals," *Bell Syst. Tech. J.*, vol. 36, pp. 653-709, May 1957.
5. K. W. Cattermole, *Principles of Pulse-Code Modulation*, Ilife, England, 1969.
6. C. L. Dammann, L. D. McDaniel, and C. L. Maddox, "D-2 Channel Bank Multiplexing and Coding," *Bell Syst. Tech. J.*, vol. 51, pp. 1675–1700, Oct. 1972.

7. Bell Telephone Laboratories, *Transmission Systems for Communication*, 4th ed., Murray Hill, NJ, 1970.
8. E. L. Gruenberg, *Handbook of Telemetry and Remote Control*, McGraw-Hill, New York, 1967.
9. J. B. O'Neal, Jr., " Delta Modulation Quantizing Noise: Analytical and Computer Simulation Results for Gaussian and Television Input Signals," *Bell Sys. Tech. J.*, pp. 117–141, Jan. 1966.
10. F. de Jager, "Delta Modulation, a Method of PCM Transmission Using the 1-Unit Code," *Philips Res. Rep.*, no. 7, pp. 442–466, 1952.
11. A. Tomozawa and H. Kaneko, "Companded Delta Modulation for Telephone Transmission," *IEEE Trans. Commun. Technol.*, vol. CT-16, pp. 149–157, Feb. 1968.
12. J. A. Greefkes and F. de Jager, "Continuous Delta Modulation," *Philips Res. Rep.*, no. 23, pp. 233–246, 1968.
13. N. S. Jayant and P. Noll, *Digital Coding of Waveform*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

PROBLEMS

6.1-1 Figure P6.1-1 shows Fourier spectra of signals $g_1(t)$ and $g_2(t)$. Determine the Nyquist interval and the sampling rate for signals $g_1(t)$, $g_2(t)$, $g_1^2(t)$, $g_2^3(t)$, and $g_1(t)g_2(t)$.



(a) (b)

Figure P6.1-1

6.1-2 Determine the Nyquist sampling rate and the Nyquist sampling interval for the signals: **(a)** sinc $(100\pi t)$; **(b)** sinc^2 $(100\pi t)$; **(c)** sinc $(100\pi t)$ + sinc $(50\pi t)$; **(d)** sinc $(100\pi t)$ + 3 sinc^2 $(60\pi t)$; **(e)** sinc $(50\pi t)$sinc $(100\pi t)$.

6.1-3 A signal $g(t)$ band-limited to B Hz is sampled by a periodic pulse train $p_{T_s}(t)$ made up of a rectangular pulse of width $1/8B$ seconds (centered at the origin) repeating at the Nyquist rate ($2B$ pulses per second). Show that the sampled signal $\overline{g}(t)$ is given by

$$\overline{g}(t) = \frac{1}{4}g(t) + \sum_{n=1}^{\infty} \frac{2}{n\pi} \sin\left(\frac{n\pi}{4}\right) g(t) \cos n\omega_s t \qquad \omega_s = 4\pi B$$

Show that the signal $g(t)$ can be recovered by passing $\overline{g}(t)$ through an ideal low-pass filter of bandwidth B Hz and a gain of 4.

6.1-4 A signal $g(t) = \text{sinc}^2(5\pi t)$ is sampled (using uniformly spaced impulses) at a rate of: **(i)** 5 Hz; **(ii)** 10 Hz; **(iii)** 20 Hz. For each of the three case:

(a) Sketch the sampled signal.

(b) Sketch the spectrum of the sampled signal.

(c) Explain whether you can recover the signal $g(t)$ from the sampled signal.

(d) If the sampled signal is passed through an ideal low-pass filter of bandwidth 5 Hz, sketch the spectrum of the output signal.

6.1-5 Signals $g_1(t) = 10^4\text{rect}\,(10^4t)$ and $g_2(t) = \delta(t)$ are applied at the inputs of ideal low-pass filters $H_1(\omega) = \text{rect}\,(\omega/40,000\pi)$ and $H_2(\omega) = \text{rect}\,(\omega/20,000\pi)$ (Fig. P6.1-5). The outputs $y_1(t)$ and $y_2(t)$ of these filters are multiplied to obtain the signal $y(t) = y_1(t)y_2(t)$. Find the Nyquist rate of $y_1(t)$, $y_2(t)$, and $y(t)$.

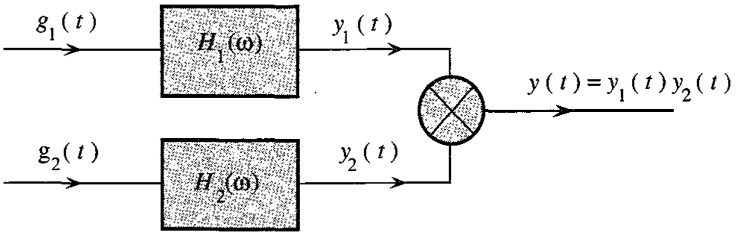


Figure P6.1-5

6.1-6 A zero-order hold circuit (Fig. P6.1-6) is often used to reconstruct a signal $g(t)$ from its samples.

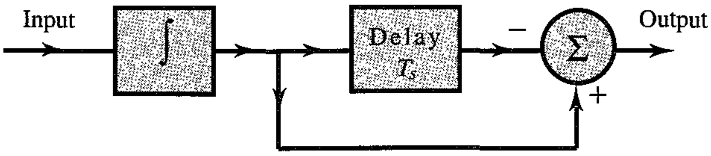


Figure P6.1-6

(a) Find the unit impulse response of this circuit.

(b) Find the transfer function $H(\omega)$ and sketch $|H(\omega)|$.

(c) Show that when a sampled signal $\overline{g}(t)$ is applied at the input of this circuit, the output is a staircase approximation of $g(t)$. The sampling interval is T_s.

6.1-7 **(a)** A first-order hold circuit can also be used to reconstruct a signal $g(t)$ from its samples. The impulse response of this circuit is

$$h(t) = \Delta\left(\frac{t}{2T_s}\right)$$

where T_s is the sampling interval. Consider a typical sampled signal $\overline{g}(t)$ and show that this circuit performs the linear interpolation. In other words, the filter output consists of sample tops connected by straight-line segments. Follow the procedure discussed in Sec. 6.1.1 (Fig. 6.2b).

(b) Determine the transfer function of this filter and its amplitude response, and compare it with the ideal filter required for signal reconstruction.

(c) This filter, being noncausal, is unrealizable. Suggest a modification that will make this filter realizable. How would such a modification affect the reconstruction of $g(t)$ from its samples? How would it affect the frequency response of the filter?

6.1-8 Prove that a signal cannot be simultaneously time-limited and band-limited. *Hint*: Show that contrary assumption leads to contradiction. Assume a signal simultaneously time-limited and band-limited so that $G(\omega) = 0$ for $|\omega| > 2\pi B$. In this case $G(\omega) = G(\omega)\ \text{rect}\,(\omega/4\pi B')$ for

$B' > B$. This means that $g(t)$ is equal to $g(t) * 2B'\text{sinc}\,(2\pi B't)$. Show that the latter cannot be time-limited.

6.2-1 The American Standard Code for Information Interchange (ASCII) has 128 characters, which are binary coded. If a certain computer generates 100,000 characters per second, determine the following:

(a) The number of bits (binary digits) required per character.

(b) The number of bits per second required to transmit the computer output, and the minimum bandwidth required to transmit this signal.

(c) For single error-detection capability, an additional bit (parity bit) is added to the code of each character. Modify your answers in parts **(a)** and **(b)** in view of this information.

6.2-2 A compact disc (CD) records audio signals digitally by using PCM. Assume the audio signal bandwidth to be 15 kHz.

(a) What is the Nyquist rate?

(b) If the Nyquist samples are quantized into $L = 65,536$ levels and then binary coded, determine the number of binary digits required to encode a sample.

(c) Determine the number of binary digits per second (bit/s) required to encode the audio signal.

(d) For practical reasons discussed in the text, signals are sampled at a rate well above the Nyquist rate. Practical CDs use 44,100 samples per second. If $L = 65,536$, determine the number of bits per second required to encode the signal, and the minimum bandwidth required to transmit the encoded signal.

6.2-3 A television signal (video and audio) has a bandwidth of 4.5 MHz. This signal is sampled, quantized, and binary coded to obtain a PCM signal.

(a) Determine the sampling rate if the signal is to be sampled at a rate 20% above the Nyquist rate.

(b) If the samples are quantized into 1024 levels, determine the number of binary pulses required to encode each sample.

(c) Determine the binary pulse rate (bits per second) of the binary-coded signal, and the minimum bandwidth required to transmit this signal.

6.2-4 Five telemetry signals, each of bandwidth 1 kHz, are to be transmitted simultaneously by binary PCM. The maximum tolerable error in sample amplitudes is 0.2% of the peak signal amplitude. The signals must be sampled at least 20% above the Nyquist rate. Framing and synchronizing requires an additional 0.5% extra bits. Determine the minimum possible data rate (bits per second) that must be transmitted, and the minimum bandwidth required to transmit this signal.

6.2-5 It is desired to set up a central station for simultaneous monitoring of the electrocardiograms (ECGs) of 10 hospital patients. The data from the rooms of the 10 patients are brought to a processing center over wires and are sampled, quantized, binary coded, and time-division multiplexed. The multiplexed data are now transmitted to the monitoring station (Fig. P6.2-5). The ECG signal bandwidth is 100 Hz. The maximum acceptable error in sample amplitudes is 0.25% of the peak signal amplitude. The sampling rate must be at least twice the Nyquist rate. Determine the minimum cable bandwidth needed to transmit these data.

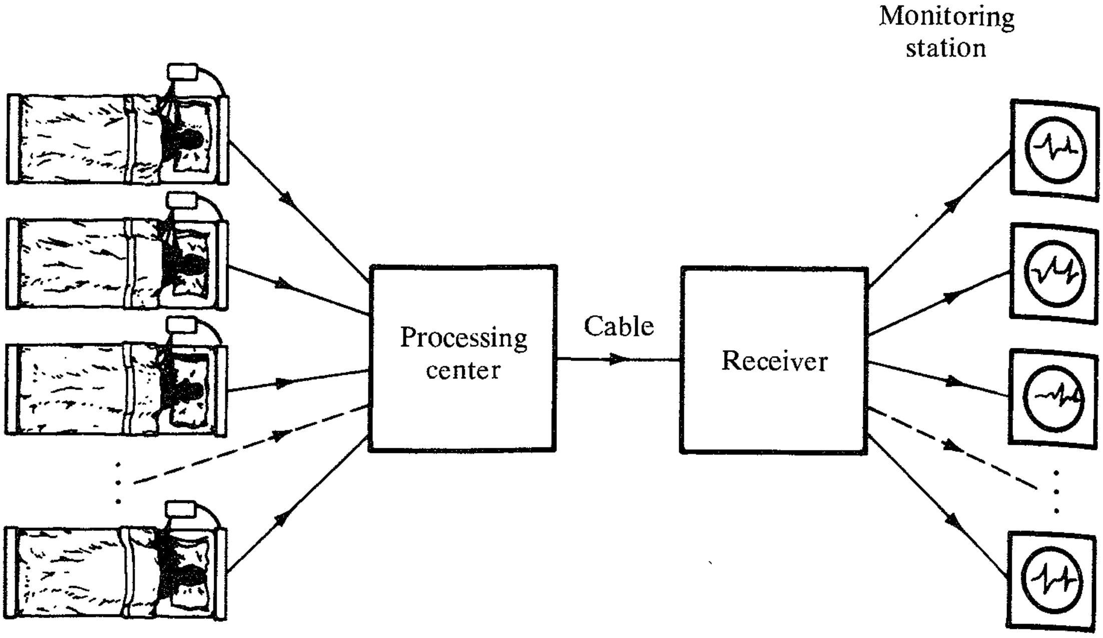


Figure P6.2-5

6.2-6 A message signal $m(t)$ is transmitted by binary PCM without compression. If the SNR (signal-to-quantization-noise ratio) is required to be at least 47 dB, determine the minimum value of L required, assuming that $m(t)$ is sinusoidal. Determine the SNR obtained with this minimum L.

6.2-7 Repeat Prob. 6.2-6 for $m(t)$ shown in Fig. P6.2-7. *Hint*: The power of a periodic signal is its energy averaged over one cycle. In this case, however, because the signal amplitude takes on the same values every quarter cycle, the power is also equal to the signal energy averaged over a quarter cycle.

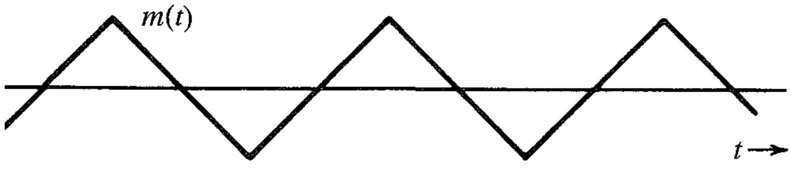


Figure P6.2-7

6.2-8 For a PCM signal, determine L if the compression parameter $\mu = 100$ and the minimum SNR required is 45 dB. Determine the output SNR with this value of L. Remember that L must be a power of 2, that is, $L = 2^n$ for a binary PCM.

6.2-9 A signal band-limited to 1 MHz is sampled at a rate 50% higher than the Nyquist rate and quantized into 256 levels using a μ-law quantizer with $\mu = 255$.

(a) Determine the signal-to-quantization-noise ratio.

(b) The SNR (the received signal quality) found in part **(a)** was unsatisfactory. It must be increased at least by 10 dB. Would you be able to obtain the desired SNR without increasing the transmission bandwidth if it was found that a sampling rate 20% above the Nyquist rate is adequate? If so, explain how. What is the maximum SNR that can be realized in this way?

6.2-10 The output SNR of a 10-bit PCM ($n = 10$) was found to be 30 dB. The desired SNR is 42 dB. It was decided to increase the SNR to the desired value by increasing the number of quantization levels L. Find the fractional increase in the transmission bandwidth required for this increase in L.

6.4-1 In a single-integration DM system, the voice signal is sampled at a rate of 64 kHz. The maximum signal amplitude $A_{\max} = 1$.

(a) Determine the minimum value of the step size σ to avoid slope overload.

(b) Determine the granular-noise power N_o if the voice signal bandwidth is 3.5 kHz.

(c) Assuming that the voice signal is sinusoidal, determine S_o and the SNR.

(d) Assuming that the voice signal amplitude is uniformly distributed in the range $(-1,\ 1)$, determine S_o and the SNR.

(e) Determine the minimum transmission bandwidth.

7 PRINCIPLES OF DIGITAL DATA TRANSMISSION

Although a significant portion of communication today is in analog form, it is being replaced rapidly by digital communication. Within the next decade most of the communication will become digital, with analog communication playing a minor role. In this chapter, we discuss various aspects of digital data transmission.

This chapter deals with the problems of transmitting digital data over a channel. Hence, the starting messages are assumed to be digital. To begin with we shall consider the binary case, where the data consists of only two symbols: **1** and **0**. We assign a distinct waveform (pulse) to each of these two symbols. The resulting sequence of these pulses is transmitted over a channel. At the receiver, these pulses are detected and are converted back to binary data (**1**s and **0**s).

7.1 A DIGITAL COMMUNICATION SYSTEM

A digital communication system is made up of several components as described in this section.

Source

The input to a digital system is in the form of a sequence of digits. The input could be the output from such sources as a data set, a computer, a digitized voice signal (PCM or DM), a digital facsimile or television, or telemetry equipment. Most of the discussion in this chapter is restricted to the binary case (communication schemes using only two symbols). A more general case of M-ary communication which uses M symbols is briefly discussed in Sec. 7.7.

Multiplexer

Generally speaking, the capacity of a practical channel transmitting data is much larger than the data rate of individual sources. To utilize this capacity effectively, we combine several sources through a digital multiplexer using the process of interleaving. Thus a channel is time-shared by several messages simultaneously.

Line Coder

The output of a multiplexer is coded into electrical pulses or waveforms for the purpose of transmission over the channel. This process is called **line coding** or **transmission coding**. There are many possible ways of assigning waveforms (pulses) to the digital data. In the binary case (two symbols), for example, conceptually the simplest line code is **on-off**, where a **1** is transmitted by a pulse $p(t)$ and a **0** is transmitted by no pulse (zero signal), as shown in Fig. 7.1a. Another commonly used code is **polar**, where **1** is transmitted by a pulse $p(t)$ and **0** is transmitted by a pulse $-p(t)$ (Fig. 7.1b). The polar scheme is the most power efficient code, because for a given noise immunity (error probability) this code requires the least power. Another popular code in PCM is **bipolar**, also known as **pseudoternary** or **alternate mark inversion (AMI)**, where **0** is encoded by no pulse and **1** is encoded by a pulse $p(t)$ or $-p(t)$, depending on whether the previous **1** is encoded by $-p(t)$ or $p(t)$. In short, pulses representing consecutive **1**'s alternate in sign, as shown in Fig. 7.1c. This code has the advantage that if an error is made in the detecting of pulses, the received pulse sequence will violate the bipolar rule and the error is immediately detected (although not corrected).*

Another line code that in the past appeared promising is the duobinary (and modified duobinary) proposed by Lender.[1, 2] Although this code is better than the bipolar in terms of

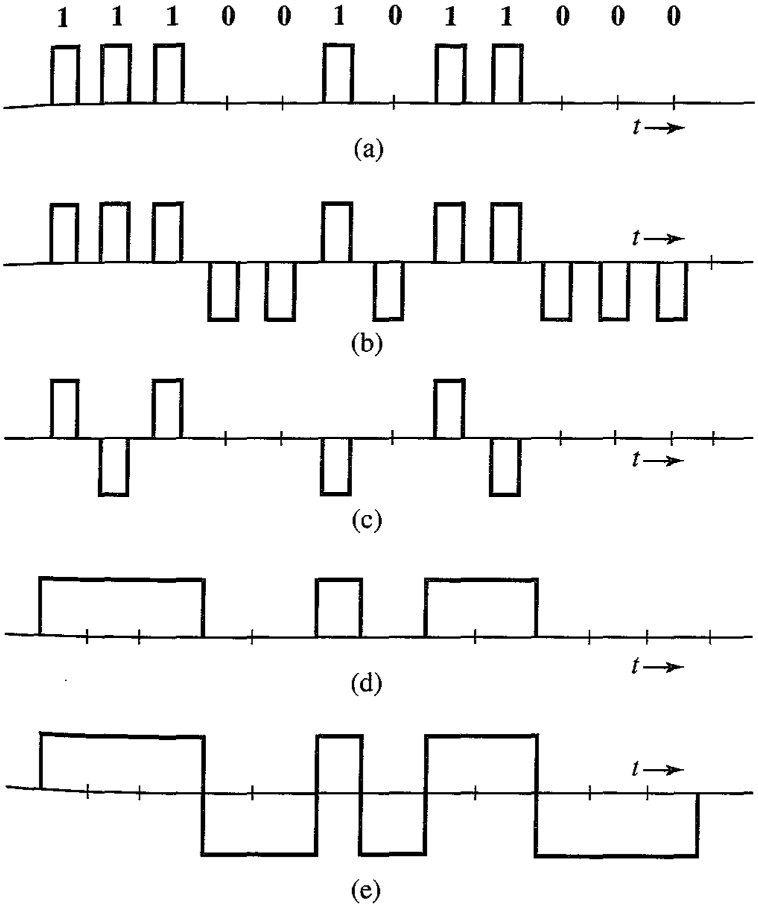


Figure 7.1 Some line codes. (a) On-off (RZ). (b) Polar (RZ). (c) Bipolar (RZ). (d) On-off (NRZ). (e) Polar (NRZ).

---

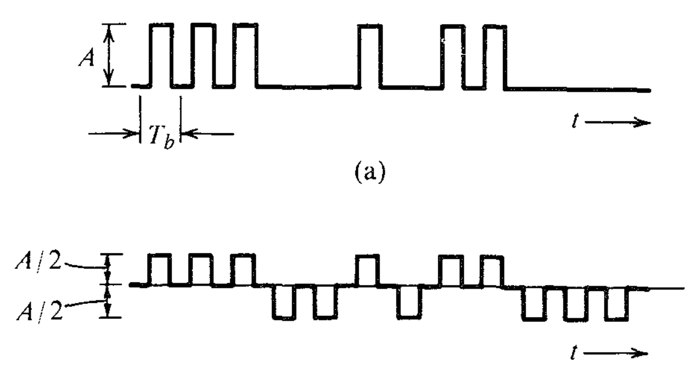
* This assumes no more than one error in sequence. Multiple errors in sequence could cancel their effects and remain undetected. However, the probability of multiple errors is much smaller than that of single errors. Even for single errors, we cannot tell exactly where the error is located. Therefore, this code can detect the presence of single errors, but cannot correct them.

bandwidth efficiency, it has lost its appeal due to some practical problems and will not be discussed here.*

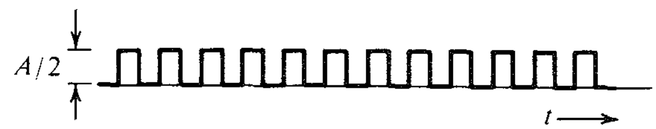
In our discussion so far, we have used half-width pulses just for the sake of illustration. We can select other widths also. Full-width pulses are often used in some applications. Whenever full-width pulses are used, the pulse amplitude is held to a constant value throughout the pulse interval (it does not have a chance to go to zero before the next pulse begins). For this reason these schemes are called **nonreturn-to-zero (NRZ)** schemes in contrast to **return-to-zero (RZ)** schemes (Fig. 7.1a, b, and c). Figure 7.1d shows an on-off NRZ signal, whereas Fig. 7.1e shows a polar NRZ signal.

Regenerative Repeater

Regenerative repeaters are used at regularly spaced intervals along a digital transmission line to detect the incoming digital signal and regenerate new clean pulses for further transmission along the line. This process periodically eliminates, and thereby combats, the accumulation of noise and signal distortion along the transmission path. If the pulses are transmitted at a rate of R_b pulses per second, we require the periodic timing information—the clock signal at R_b Hz—to sample the incoming pulses at a repeater. This timing information can be extracted from the received signal itself if the line code is chosen properly. The polar signal in Fig. 7.1b, for example, when rectified, results in a periodic signal of clock frequency R_b Hz, which contains the desired periodic timing signal of frequency R_b Hz. When this signal is applied to a resonant circuit tuned to frequency R_b, the output, which is a sinusoid of frequency R_b Hz, can be used for timing. The on-off signal can be expressed as the sum of a periodic signal (of clock frequency) and a polar signal, as shown in Fig. 7.2. Because of the presence of the



(b) Random component



(c) Periodic component

Figure 7.2 An on-off signal is the sum of a polar signal and a clock frequency periodic signal.

---

* Modified duobinary has seen some limited application in the North American digital telephone network. A reversal between the two conductors of a cable pair (tip and ring) in splicing operations will not affect a bipolar signal since positive and negative pulses have the same logic value. Nor will it adversely affect most of the other types of services found in a telephone trunk plant. It will, however, cause the failure of the polarity-sensitive modified duobinary system. The presence of numerous "tip-ring turnovers" in North American trunk cable plant, along with the advent of lightwave facilities as a replacement for paired cable systems, prevented duobinary from displacing bipolar transmission in terrestrial applications, except in certain special cases.

periodic component, we can extract the timing information from this signal using a resonant circuit tuned to the clock frequency. A bipolar signal, when rectified, becomes an on-off signal. Hence, its timing information can be extracted the same way as that for an on-off signal.

The timing signal (the resonant circuit output) is sensitive to the incoming bit pattern. In the on-off or bipolar case, a **0** is transmitted by "no pulse." Hence, if there are too many **0**'s in a sequence (no pulses), there is no signal at the input of the resonant circuit and the sinusoidal output of the resonant circuit starts decaying, thus causing error in the timing information. We shall discuss later the ways of overcoming this problem. A line code in which the bit pattern does not affect the accuracy of the timing information is said to be a **transparent** line code. The polar scheme (where each bit is transmitted by some pulse) is transparent, whereas the on-off and bipolar schemes are nontransparent.

7.2 LINE CODING

Digital data can be transmitted by various **transmission** or **line codes**, such as on-off, polar, bipolar, and so on. Each has its advantages and disadvantages. Among other desirable properties, a line code should have the following properties:

1. *Transmission bandwidth*: It should be as small as possible.
2. *Power efficiency*: For a given bandwidth and a specified detection error probability, the transmitted power should be as small as possible.
3. *Error detection and correction capability*: It should be possible to detect, and preferably correct, detection errors. In a bipolar case, for example, a single error will cause bipolar violation and can easily be detected. Error-correcting codes will be discussed in depth in Chapter 16.*
4. *Favorable power spectral density*: It is desirable to have zero PSD at $\omega = 0$ (dc), because ac coupling and transformers are used at the repeaters. Significant power in low-frequency components causes dc wander in the pulse stream when ac coupling is used. The ac coupling is required because the dc paths provided by the cable pairs between the repeater sites are used to transmit the power required to operate the repeaters.
5. *Adequate timing content*: It should be possible to extract timing or clock information from the signal.
6. *Transparency*: It should be possible to transmit a digital signal correctly regardless of the pattern of **1**'s and **0**'s. We saw earlier that a long string of **0**'s could cause errors in timing extraction in on-off and bipolar cases. If the data are so coded that for every possible sequence of data the coded signal is received faithfully, the code is transparent.

PSD of Various Line Codes

In Example 3.23 we discussed a procedure for finding the PSD of a polar pulse train. We shall use a similar procedure to find a general expression for PSD of a large class of line codes.

---

* Since multiplexing strips out bipolar violations (BPV), which are required for such error detection, signal formats such as the extended superframe (ESF) format are displacing older formats to allow the detection of errors without resorting to an analysis of the bipolar condition of the signal.

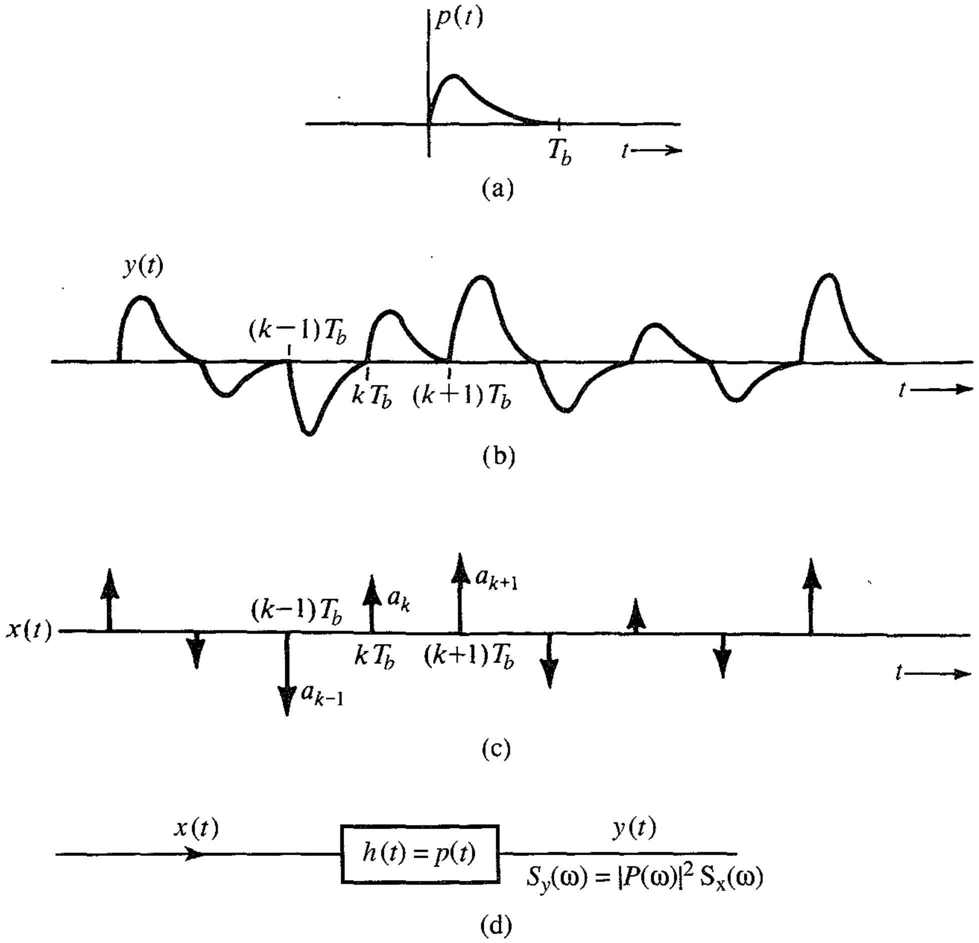


Figure 7.3 A random PAM signal and its generation from a PAM impulse sequence.

In the following discussion, the pulses are spaced T_b seconds apart. Consequently, the transmission rate is $R_b = 1/T_b$ pulses per second. The basic pulse used is denoted by $p(t)$ and its Fourier transform is $P(\omega)$.

Consider the pulse train in Fig. 7.3b constructed from a basic pulse $p(t)$ (Fig. 7.3a) repeating at intervals of T_b with relative strength a_k for the pulse starting at $t = kT_b$. In other words, the kth pulse in this pulse train $y(t)$ is $a_k p(t)$. The values a_k are arbitrary and random. This is a PAM signal. The on-off, polar, and bipolar line codes are all special cases of this pulse train $y(t)$, where a_k takes on values 0, 1, or -1 randomly subject to some constraints. We can, therefore, analyze many line codes from the knowledge of the PSD of $y(t)$. Unfortunately, it suffers from the disadvantage that it is restricted to only certain pulse shapes $p(t)$. If the pulse shape changes, we have to derive the PSD all over again. This difficulty can be overcome by a simple artifice of considering a PAM signal $x(t)$ that uses a unit impulse for the basic pulse $p(t)$ (Fig. 7.3c). The impulses are at the intervals of T_b and the strength (area) of the kth impulse is a_k. If $x(t)$ is applied to the input of a filter that has a unit impulse response $h(t) = p(t)$ (Fig. 7.3d), the output will be the pulse train $y(t)$ in Fig. 7.3b. Also, $S_y(\omega)$, the PSD of $y(t)$, is $|P(\omega)|^2 S_x(\omega)$ [see Eq. (3.90)]. This equation allows us to determine $S_y(\omega)$, the PSD of a line code corresponding to any pulse shape $p(t)$, once we know $S_x(\omega)$. This approach is attractive because of its generality. We now need to derive $\mathcal{R}_x(\tau)$, the time-autocorrelation function of the impulse train $x(t)$. This can be conveniently done by considering the impulses as a limiting form of the rectangular pulses, as shown in Fig. 7.4a. Each pulse has a width $\epsilon \to 0$ and the kth pulse height is h_k. Because the strength of the kth impulse is a_k,

$$\epsilon h_k = a_k$$

If we designate the corresponding rectangular pulse train by $\hat{x}(t)$, then by definition [Eq. (3.82b)],

$$\mathcal{R}_{\hat{x}}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \hat{x}(t)\hat{x}(t-\tau)\,dt \tag{7.1}$$

Because $\mathcal{R}_{\hat{x}}(\tau)$ is an even function of τ [Eq. (3.83)], we need consider only positive τ. To begin with, consider the case of $\tau < \epsilon$. In this case the integral in Eq. (7.1) is the area under the signal $\hat{x}(t)$ multiplied by $\hat{x}(t)$ delayed by τ $(\tau < \epsilon)$. As seen from Fig. 7.4b, the area associated with the kth pulse is $h_k^2(\epsilon - \tau)$, and

$$\begin{aligned}\mathcal{R}_{\hat{x}} &= \lim_{T\to\infty} \frac{1}{T} \sum_k h_k^2(\epsilon-\tau) \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_k a_k^2 \left(\frac{\epsilon-\tau}{\epsilon^2}\right) \\ &= \frac{R_0}{\epsilon T_b}\left(1-\frac{\tau}{\epsilon}\right) \qquad (7.2a)\end{aligned}$$

where

$$R_0 = \lim_{T\to\infty} \frac{T_b}{T} \sum_k a_k^2 \tag{7.2b}$$

During the averaging interval T $(T \to \infty)$ there are N pulses $(N \to \infty)$, where

$$N = \frac{T}{T_b} \tag{7.3}$$

and from Eq. (7.2b),

$$R_0 = \lim_{N\to\infty} \frac{1}{N} \sum_k a_k^2 \tag{7.4}$$

Observe that the summation is over N pulses. Hence, R_0 is the time average of the square of the pulse amplitudes a_k. Using our time average notation, we can express R_0 as

$$R_0 = \lim_{N\to\infty} \frac{1}{N} \sum_k a_k^2 = \widetilde{a_k^2} \tag{7.5}$$

We also know that $\mathcal{R}_{\hat{x}}(\tau)$ is an even function of τ [see Eq. (3.83)]. Hence, Eq. (7.2a) can be expressed as

$$\mathcal{R}_{\hat{x}}(\tau) = \frac{R_0}{\epsilon T_b}\left(1-\frac{|\tau|}{\epsilon}\right) \qquad |\tau| < \epsilon \tag{7.6}$$

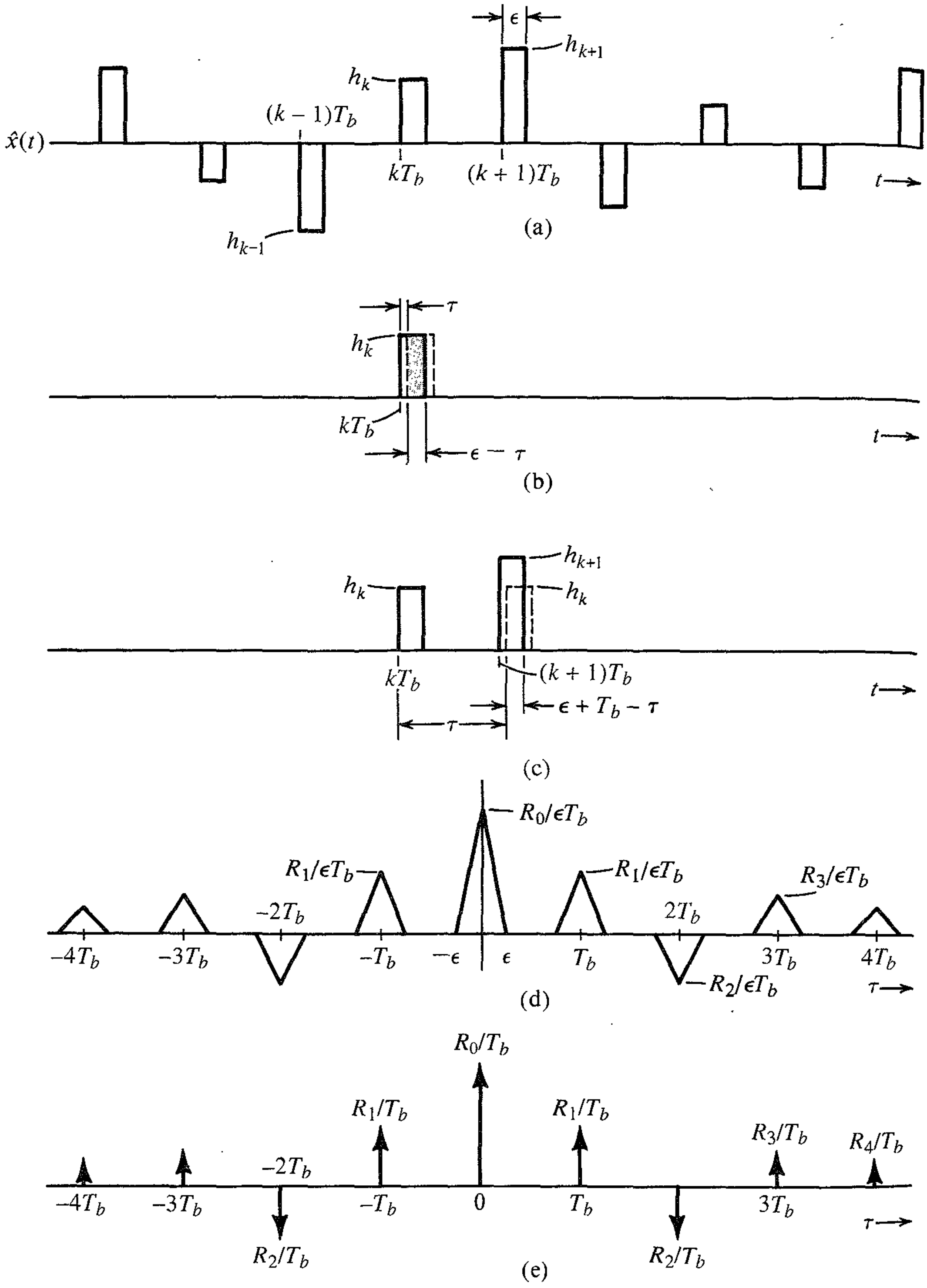

$\hat{x}(t)$
ϵ
h_{k+1}
h_k
$(k-1)T_b$
kT_b
$(k+1)T_b$
h_{k-1}
$t \rightarrow$
(a)
τ
h_k
kT_b
$\epsilon - \tau$
$t \rightarrow$
(b)
h_{k+1}
h_k
h_k
kT_b
$(k+1)T_b$
$\epsilon + T_b - \tau$
τ
$t \rightarrow$
(c)
$R_0/\epsilon T_b$
$R_1/\epsilon T_b$
$R_1/\epsilon T_b$
$R_3/\epsilon T_b$
$-2T_b$
$2T_b$
$-4T_b$
$-3T_b$
$-T_b$
$-\epsilon$
ϵ
T_b
$3T_b$
$4T_b$
$R_2/\epsilon T_b$
$\tau \rightarrow$
(d)
R_0/T_b
R_1/T_b
R_1/T_b
R_3/T_b
R_4/T_b
$-2T_b$
$-4T_b$
$-3T_b$
$-T_b$
0
T_b
$3T_b$
$\tau \rightarrow$
R_2/T_b
R_2/T_b
(e)

Figure 7.4 Deviation of PSD of a random binary signal.

This is a triangular pulse of height $R_0/\epsilon T_b$ and width 2ϵ centered at $\tau = 0$ (Fig. 7.4d). This is expected because as τ increases beyond ϵ, there is no overlap between the delayed signal $\hat{x}(t-\tau)$ and $\hat{x}(t)$, and hence, $\mathcal{R}_{\hat{x}}(\tau) = 0$ as seen from Fig. 7.4d. But as we increase τ further, we find that the kth pulse of $\hat{x}(t-\tau)$ will start overlapping the $(k+1)$th pulse of $\hat{x}(t)$ as τ approaches T_b (Fig. 7.4c). Repeating the earlier argument, we see that $\mathcal{R}_{\hat{x}}(\tau)$ will have another triangular pulse of width 2ϵ centered at $\tau = T_b$ and of height $R_1/\epsilon T_b$, where

$$\begin{aligned} R_1 &= \lim_{T\to\infty} \frac{T_b}{T} \sum_k a_k a_{k+1} \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_k a_k a_{k+1} \\ &= \widetilde{a_k a_{k+1}} \end{aligned}$$

Observe that R_1 is obtained by multiplying every pulse strength a_k with the strength of its immediate neighbor a_{k+1}, adding all these products, and then dividing the sum by the total number of pulses. This is clearly the time average (mean) of the product $a_k a_{k+1}$ and is, by our notation, $\widetilde{a_k a_{k+1}}$. A similar thing happens around $\tau = 2T_b,\ 3T_b,\ \ldots$. Hence, $\mathcal{R}_{\hat{x}}(\tau)$ consists of a sequence of triangular pulses of width 2ϵ centered at $\tau = 0,\ \pm T_b,\ \pm 2T_b,\ \ldots$. The height of the pulses centered at $\pm nT_b$ is $R_n/\epsilon T_b$, where

$$\begin{aligned} R_n &= \lim_{T\to\infty} \frac{T_b}{T} \sum_k a_k a_{k+n} \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_k a_k a_{k+n} \\ &= \widetilde{a_k a_{k+n}} \end{aligned}$$

In order to find $\mathcal{R}_x(\tau)$, we let $\epsilon \to 0$ in $\mathcal{R}_{\hat{x}}(\tau)$. As $\epsilon \to 0$, the width of each triangular pulse $\to 0$ and the height $\to \infty$ in such a way that the area is still finite. Thus, in the limit as $\epsilon \to 0$, the triangular pulses become impulses. For the nth pulse centered at nT_b, the height is $R_n/\epsilon T_b$ and the area is R_n/T_b. Hence (Fig. 7.4e),

$$\mathcal{R}_x(\tau) = \frac{1}{T_b} \sum_{n=-\infty}^{\infty} R_n\, \delta(\tau - nT_b) \tag{7.7}$$

The PSD $S_x(\omega)$ is the Fourier transform of $\mathcal{R}_x(\tau)$. Therefore,

$$S_x(\omega) = \frac{1}{T_b} \sum_{n=-\infty}^{\infty} R_n e^{-jn\omega T_b} \tag{7.8}$$

Recognizing the fact that $R_{-n} = R_n$ [because $\mathcal{R}(\tau)$ is an even function of τ], we have

$$S_x(\omega) = \frac{1}{T_b} \left(R_0 + 2 \sum_{n=1}^{\infty} R_n \cos n\omega T_b \right) \tag{7.9}$$

The input $x(t)$ to the filter with impulse response $h(t) = p(t)$ results in the output $y(t)$, as shown in Fig. 7.3d. If $p(t) \Longleftrightarrow P(\omega)$, the transfer function of the filter is $H(\omega) = P(\omega)$, and according to Eq. (3.90),

$$S_y(\omega) = |P(\omega)|^2 S_x(\omega) \tag{7.10a}$$

$$= \frac{|P(\omega)|^2}{T_b}\left(\sum_{n=-\infty}^{\infty} R_n e^{-jn\omega T_b}\right) \tag{7.10b}$$

$$= \frac{|P(\omega)|^2}{T_b}\left(R_0 + 2\sum_{n=1}^{\infty} R_n \cos n\omega T_b\right) \tag{7.10c}$$

Using this result, we shall now find the PSDs of various line codes.

Polar Signaling

In polar signaling, **1** is transmitted by a pulse $p(t)$ and **0** is transmitted by $-p(t)$. In this case, a_k is equally likely to be 1 or -1, and a_k^2 is always 1. Hence,

$$R_0 = \lim_{N\to\infty} \frac{1}{N} \sum_k a_k^2$$

There are N pulses and $a_k^2 = 1$ for each one. The summation on the right-hand side of this equation is N. Hence,

$$R_0 = \lim_{N\to\infty} \frac{1}{N}(N) = 1 \tag{7.11a}$$

Moreover, both a_k and a_{k+1} are either 1 or -1. Hence, $a_k a_{k+1}$ is either 1 or -1. Because the pulse amplitude a_k is equally likely to be 1 and -1 on the average, out of N terms the product $a_k a_{k+1}$ is equal to 1 for $N/2$ terms and is equal to -1 for the remaining $N/2$ terms. Therefore,

$$R_1 = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{2}(1) + \frac{N}{2}(-1)\right] = 0 \tag{7.11b}$$

Arguing this way, we see that the product $a_k a_{k+n}$ is also equally likely to be 1 or -1. Hence,

$$R_n = 0 \qquad n \geq 1 \tag{7.11c}$$

Therefore from Eq. (7.10c),

$$S_y(\omega) = \frac{|P(\omega)|^2}{T_b} R_0$$

$$= \frac{|P(\omega)|^2}{T_b} \tag{7.12}$$

For the sake of comparison of various schemes, we shall consider a specific pulse shape. Let $p(t)$ be a rectangular pulse of width $T_b/2$ (a half-width rectangular pulse), that is,

$$p(t) = \text{rect}\left(\frac{t}{T_b/2}\right) = \text{rect}\left(\frac{2t}{T_b}\right)$$

and

$$P(\omega) = \frac{T_b}{2}\,\text{sinc}\left(\frac{\omega T_b}{4}\right) \tag{7.13}$$

Therefore,

$$S_y(\omega) = \frac{T_b}{4}\,\text{sinc}^2\left(\frac{\omega T_b}{4}\right) \tag{7.14}$$

Figure 7.5 shows the spectrum $S_y(\omega)$. From this spectrum, the essential bandwidth of the signal is seen to be $2R_b$ Hz (where R_b is the clock frequency). This is four times the theoretical bandwidth (Nyquist bandwidth) required to transmit R_b pulses per second. Increasing the pulse width reduces the bandwidth (expansion in the time domain results in compression in the frequency domain). For a full-width pulse* (maximum possible pulse width), the essential bandwidth is half, that is, R_b Hz. This is still twice the theoretical bandwidth. Thus, polar signaling is not bandwidth efficient.

Second, polar signaling has no error-detection or error-correction capability. A third disadvantage of polar signaling is that it has nonzero PSD at dc ($\omega = 0$). This will rule out the use of ac coupling during transmission. The ac coupling, which permits transformers and blocking capacitors to aid in impedance matching and bias removal, and which allows dc powering of the line repeaters over the cable pairs, is very important in practice. Later, we shall show how a PSD of a line code may be forced to zero at dc by properly shaping $p(t)$.

On the positive side, polar signaling is the most efficient scheme from the power requirement viewpoint. For a given power, it can be shown that the detection-error probability for a polar scheme is the smallest possible (see Sec. 7.6). Polar signaling is also transparent because there is always some pulse (positive or negative) regardless of the bit sequence. There is no discrete clock frequency component in the spectrum of the polar signal. Rectification of the polar signal, however, yields a periodic signal of the clock frequency and can readily be used to extract timing.

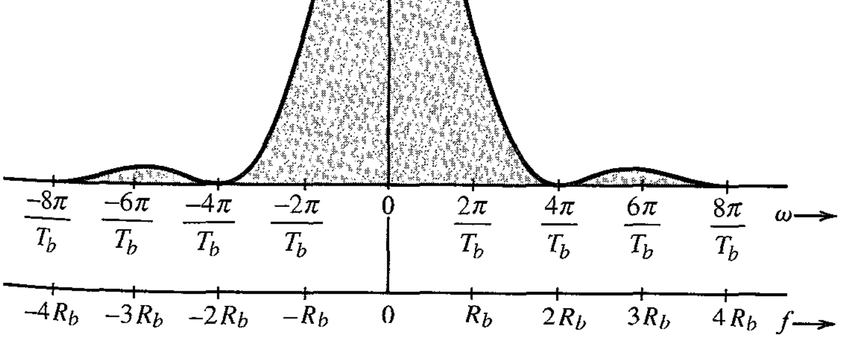


Figure 7.5 Power spectral density of a polar signal.

---

* The scheme using the full-width pulse $p(t) = \text{rect}\,(t/T_b)$ is an example of an NRZ scheme. The half-width pulse scheme, on the other hand, is an example of an RZ scheme.

Achieving a DC Null in PSD by Pulse Shaping

Because $S_y(\omega)$, the PSD of a line code, contains a factor $|P(\omega)|^2$, we can force the PSD to have a dc null by selecting a pulse $p(t)$ such that $P(\omega)$ is zero at dc ($\omega = 0$). Because

$$P(\omega) = \int_{-\infty}^{\infty} p(t)e^{-j\omega t}\, dt$$

we have

$$P(0) = \int_{-\infty}^{\infty} p(t)\, dt$$

Hence, if the area under $p(t)$ is made zero, $P(0)$ is zero, and we have a dc null in the PSD. For a rectangular pulse, one possible shape of $p(t)$ to accomplish this is shown in Fig. 7.6a. When we use this pulse with polar line coding, the resulting signal is known as **Manchester**, or **split-phase** (also **twinned-binary**) **signal**. Using Eq. (7.12), the reader can show that for this pulse, the PSD of the Manchester line code has a dc null (see Prob. 7.2-2).

On-Off Signaling

In this case a **1** is transmitted by a pulse $p(t)$ and a **0** is transmitted by no pulse. Hence, a pulse strength a_k is equally likely to be 1 or 0. Out of N pulses in the interval of T seconds, a_k is 1 for $N/2$ pulses and 0 for the remaining $N/2$ pulses on the average. Hence,

$$R_0 = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{2}(1) + \frac{N}{2}(0)\right] = \frac{1}{2} \tag{7.15}$$

To compute R_n we need to consider the product $a_k a_{k+n}$. Since a_k and a_{k+n} are equally likely to be 1 or 0, the product $a_k a_{k+n}$ is equally likely to be 1×1, 1×0, 0×1, or 0×0, that is, 1, 0, 0, 0. Therefore, on the average, the product $a_k a_{k+n}$ is equal to 1 for $N/4$ terms and 0 for $3N/4$ terms, and

$$R_n = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{4}(1) + \frac{3N}{4}(0)\right] = \frac{1}{4} \qquad n \geq 1 \tag{7.16}$$

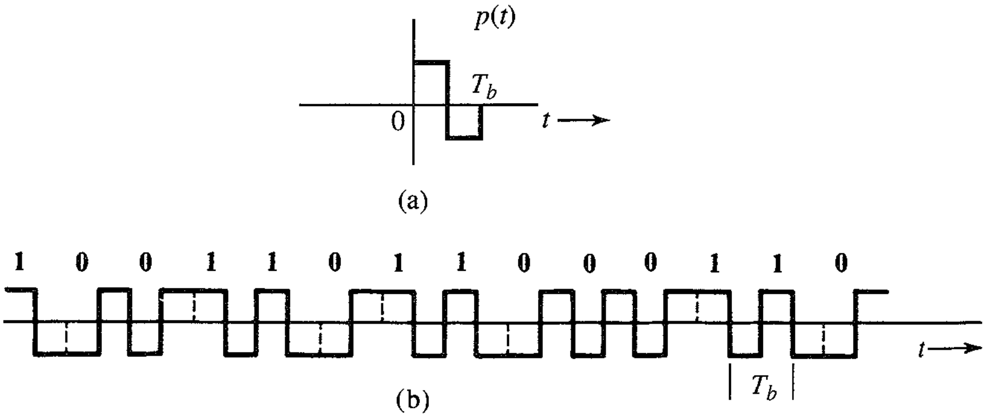


Figure 7.6 Split-phase (Manchester or twinned-binary) signal. (a) Basic pulse $p(t)$ for Manchester signaling. (b) PSD of Manchester signaling.

Therefore [Eq. (7.8)],

$$S_x(\omega) = \frac{1}{2T_b} + \frac{1}{4T_b} \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} e^{-jn\omega T_b} \tag{7.17a}$$

$$= \frac{1}{4T_b} + \frac{1}{4T_b} \sum_{n=-\infty}^{\infty} e^{-jn\omega T_b} \tag{7.17b}$$

Equation (7.17b) is obtained from Eq. (7.17a) by splitting the term $1/2T_b$ corresponding to R_0 into two: $1/4T_b$ outside the summation and $1/4T_b$ inside the summation (corresponding to $n = 0$). We now use the formula (see the footnote for a proof)*

$$\sum_{n=-\infty}^{\infty} e^{-jn\omega T_b} = \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)$$

Substitution of this result into Eq. (7.17b) yields

$$S_x(\omega) = \frac{1}{4T_b} + \frac{2\pi}{4T_b^2} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right) \tag{7.18a}$$

and the desired PSD of the on-off waveform $y(t)$ is [Eq. (7.10a)]

$$S_y(\omega) = \frac{|P(\omega)|^2}{4T_b}\left[1 + \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)\right] \tag{7.18b}$$

For the case of a half-width rectangular pulse [see Eq. (7.13)]

$$S_y(\omega) = \frac{T_b}{16}\,\text{sinc}^2\left(\frac{\omega T_b}{4}\right)\left[1 + \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)\right] \tag{7.19}$$

Note that the spectrum (Fig. 7.7) consists of both a discrete and a continuous part. A discrete component of clock frequency ($R_b = 1/T_b$) is present in the spectrum. The continuous component of the spectrum is $(T_b/16)\,\text{sinc}^2(\omega T_b/4)$. This is identical (except for a scaling factor) to the spectrum of the polar signal [Eq. (7.14)]. This is a logical result because as Fig. 7.2 shows, an on-off signal can be expressed as the sum of a polar and a periodic component. The polar component $y_1(t)$ is exactly half the polar signal discussed earlier. Hence, the PSD of this

---

* The impulse train in Fig. (3.24a) with $T_0 = T_b$ is $\delta_{T_b}(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_b)$. Moreover, the Fourier series for this impulse train as found in Eq. (2.89) is

$$\sum_{n=-\infty}^{\infty} \delta(t - nT_b) = \frac{1}{T_b} \sum_{n=-\infty}^{\infty} e^{-jn\omega_b t} \qquad \omega_b = \frac{2\pi}{T_b}$$

We take the Fourier transform of both sides of this equation and use the fact that $\delta(t - nT_b) \Longleftrightarrow e^{-jn\omega T_b}$ and $e^{-jn\omega_b t} \Longleftrightarrow 2\pi\delta(\omega - \omega_b)$. This yields

$$\sum_{n=-\infty}^{\infty} e^{-jn\omega T_b} = \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)$$

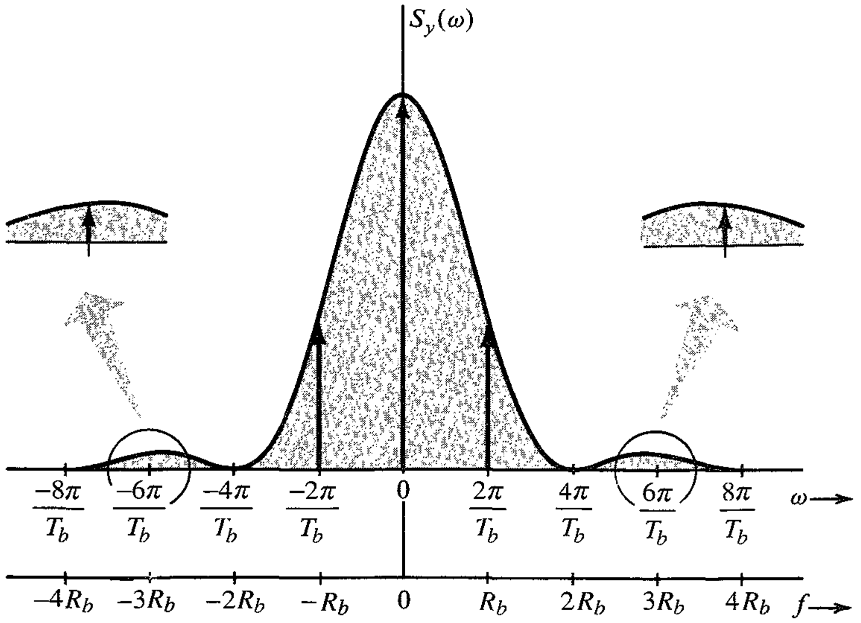


Figure 7.7 Power spectral density of an on-off signal.

component is one-fourth the PSD in Eq. (7.14). The periodic component is of clock frequency R_b, and consists of discrete components of frequency R_b and its harmonics.

There is very little to recommend in on-off signaling. For a given transmitted power, it is less immune to noise interference than the polar scheme, which uses a positive pulse for **1** and a negative pulse for **0**. This is because the noise immunity depends on the difference of amplitudes representing **1** and **0**. Hence, for the same immunity, if on-off signaling uses pulses of amplitudes 2 and 0, polar signaling need only use pulses of amplitudes 1 and -1. It is simple to show that on-off signaling requires twice as much power as polar signaling. If a pulse of amplitude 1 or -1 has energy E, then the pulse of amplitude 2 has energy $(2)^2E = 4E$. Because $1/T_b$ digits are transmitted per second, polar signal power is $E(1/T_b) = E/T_b$. For the on-off case, on the other hand, each pulse energy is $4E$, but only half as many pulses are transmitted. Hence, the signal power is $4E(1/2T_b) = 2E/T_b$, which is twice that required for the polar signal. Moreover, unlike the polar case, on-off signaling is not transparent. A long string of **0**'s (or offs) causes the absence of a signal and can create an error in timing extraction. In addition, all the disadvantages of polar signaling [e.g., excessive transmission bandwidth, nonzero power spectrum at dc, no error detection (or correction) capability] are also present in on-off signaling.

Bipolar [Pseudoternary or Alternate Mark Inverted (AMI)] Signaling

This is the signaling scheme used in PCM these days. A **0** is transmitted by no pulse, and a **1** is transmitted by a pulse $p(t)$ or $-p(t)$, depending on whether the previous **1** was transmitted by $-p(t)$ or $p(t)$. With consecutive pulses alternating, we can avoid the dc wander and thus cause a dc null in the PSD. Bipolar signaling actually uses three symbols [$p(t)$, 0, and $-p(t)$], and, hence, it is in reality ternary rather than binary signaling.

To calculate the PSD, we have

$$R_0 = \lim_{N\to\infty} \frac{1}{N} \sum_k a_k^2$$

On the average, half of the a_k's are 0, and the remaining half are either 1 or -1, with $a_k^2 = 1$. Therefore,

$$R_0 = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{2}(\pm 1)^2 + \frac{N}{2}(0)\right] = \frac{1}{2}$$

To compute R_1, we consider the pulse strength product $a_k a_{k+1}$. There are four possible equally likely sequences of two bits: **11, 10, 01, 00**. Since bit **0** is encoded by no pulse ($a_k = 0$), the product $a_k a_{k+1} = 0$ for the last three of these sequences. This means that, on the average, $3N/4$ combinations have $a_k a_{k+1} = 0$ and only $N/4$ combinations have nonzero $a_k a_{k+1}$. Because of the bipolar rule, the bit sequence **11** can only be encoded by two consecutive pulses of opposite polarities. This means the product $a_k a_{k+1} = -1$ for the $N/4$ combinations. Therefore,

$$R_1 = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{4}(-1) + \frac{3N}{4}(0)\right] = -\frac{1}{4}$$

To compute R_2 in a similar way, we need to observe the product $a_k a_{k+2}$. For this, we need to consider all possible combinations of three bits in sequence. There are eight equally likely combinations: **111, 101, 110, 100, 011, 010, 001, 000**. The last six combinations have either the first or the last bit **0**, or both. Hence, $a_k a_{k+2} = 0$ for all these six combinations. The first two combinations are the only ones that yield nonzero $a_k a_{k+2}$. Using the bipolar rule, the first and the third pulses in the combination **111** are of the same polarity, yielding $a_k a_{k+2} = 1$. But for **101**, the first and third pulses are of opposite polarity, yielding $a_k a_{k+2} = -1$. Thus, on the average, $a_k a_{k+2} = 1$ for $N/8$ terms, -1 for $N/8$ terms, and 0 for $3N/4$ terms. Hence,

$$R_2 = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{8}(1) + \frac{N}{8}(-1) + \frac{3N}{8}(0)\right] = 0$$

In general,

$$R_n = \lim_{N\to\infty} \frac{1}{N}\sum_k a_k a_{k+n}$$

For $n > 1$, the product $a_k a_{k+2}$ can be 1, -1, or 0. Moreover, an equal number of combinations have values 1 and -1. This causes $R_n = 0$. Thus,

$$R_n = 0 \qquad n > 1$$

and [see Eq. (7.10c)]

$$S_y(\omega) = \frac{|P(\omega)|^2}{2T_b}\left[1 - \cos\omega T_b\right] \tag{7.20a}$$

$$= \frac{|P(\omega)|^2}{T_b}\sin^2\left(\frac{\omega T_b}{2}\right) \tag{7.20b}$$

Note that $S_y(\omega) = 0$ for $\omega = 0$ (dc), regardless of $P(\omega)$. Hence, the PSD has a dc null, which is desirable for ac coupling. Moreover, $\sin^2(\omega T_b/2) = 0$ at $\omega = 2\pi/T_b$, that is, at $1/T_b = R_b$ Hz. Thus, regardless of $P(\omega)$, we are assured of a bandwidth of R_b Hz. For the half-width pulse,

$$S_y(\omega) = \frac{T_b}{4}\,\text{sinc}^2\left(\frac{\omega T_b}{4}\right)\sin^2\left(\frac{\omega T_b}{2}\right) \tag{7.21}$$

This is shown in Fig. 7.8. The essential bandwidth of the signal is R_b ($R_b = 1/T_b$), which is l that of polar or on-off signaling and twice the theoretical minimum bandwidth. Observe t we were able to obtain the bandwidth R_b for the polar (or on-off) case for the full-width pul For the bipolar case, the bandwidth is R_b Hz, regardless of whether the pulse is half-width full-width.

Bipolar signaling has several advantages: (1) Its spectrum has a dc null. (2) Its bandwic is not excessive. (3) It has single-error-detection capability. This is due to the fact that i single detection error is made, it will cause a violation of the alternating pulse rule and will detected immediately. If a bipolar signal is rectified, we get an on-off signal that has a discre component at the clock frequency. Among the disadvantages, a bipolar signal requires twice much power (3 dB) as a polar signal. This is because bipolar detection is essentially equivale to on-off signaling from the detection point of view. One has to distinguish between $+A$ $-A$ and 0 rather than between $-A/2$ and $A/2$.

Another disadvantage of bipolar signaling is that it is not transparent. In practice, variou substitution schemes are used to prevent long strings of logical zeros from allowing th extracted clock signals to decay away. We shall now discuss two such schemes: high-densit bipolar (HDB) signaling and binary with 8 zero substitution (B8ZS) signaling.

High-Density Bipolar (HDB) Signaling: In this scheme, the problem of the bipola signal being nontransparent is eliminated by adding pulses when the number of consecutive **0**'s exceeds n. Such a modified coding is called **high-density bipolar (HDB) coding** and denoted by HDBN, where N can take on any value 1, 2, 3, ... The most important of the HDB codes is the HDB3 format, which has been adopted as an international standard.

The basic idea of the HDBN code is that when a run of $N + 1$ zeros occurs, this group of zeros is replaced by one of the special $N + 1$ binary digit sequences. The sequences are chosen to include some binary **1**'s in order to increase the timing content of the signal. The **1**'s included deliberately violate the bipolar rule for easy identification of the substituted sequence. In HDB3 coding, for example, the special sequences used are **000V** and **B00V**, where **B=1** conforms to the bipolar rule and **V=1** violates the bipolar rule. The choice of sequence **000V**

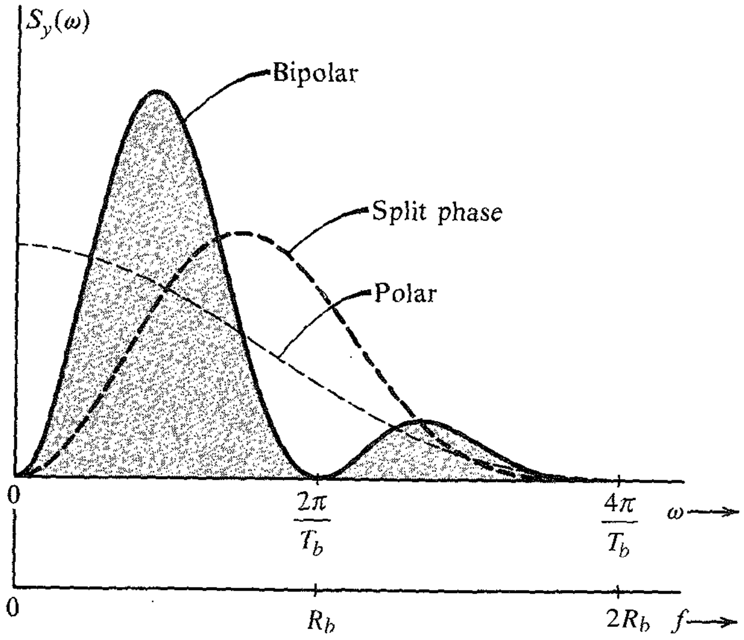


Figure 7.8 PSD of bipolar, polar, and split-phase signals normalized for equal powers. Half-width rectangular pulses are used.

or **B00V** is made in such a way that consecutive **V** pulses alternate signs in order to avoid dc wander and to maintain the dc null in the PSD. This requires that the sequence **B00V** be used when there is an even number of **1**'s following the last special sequence and the sequence **000V** be used when there is an odd number of **1**'s following the last sequence. Figure 7.9a shows an example of this coding. Note that in the sequence **B00V**, **B** and **V** are both encoded by the same pulse. The decoder has to check two things—the bipolar violations and the number of **0**'s preceding each violation to determine if the previous **1** is also a substitution.

Despite deliberate bipolar violations, HDB signaling retains error-detecting capability. Any single error will insert a spurious bipolar violation (or will delete one of the deliberate violations). This will become apparent when, at the next violation, the alternation of violations does not appear. This also shows that deliberate violations can be detected despite single errors. Figure 7.9b shows the PSD of HDB3 as well as that of a bipolar signal to facilitate comparison[3].

Binary with 8 Zero Substitution (B8ZS) Signaling: A class of line codes similar to HDBN is the **BNZS code**, where if N zeros occur in succession, they are substituted by one of the two special sequences containing some **1**'s to increase timing content. There are deliberate bipolar violations just as in HDBN. **Binary with eight zero substitution (B8ZS)**, is such a scheme used in DS1 signals, It replaces any string of eight zeros in length with a sequence of **1**'s and **0**'s containing two bipolar violations. Such a sequence is unlikely to be counterfeited by errors, and any such sequence received by a digital channel bank is replaced by a string of eight logical zeros prior to decoding. The sequence used as a replacement consists of the pattern **000VB0VB**. Similarly, in B6ZS code used in DS2 signals, a string of six zeros is replaced with **0VB0VB**, and DS3 signal features a three-zero B3ZS code. The B3ZS code is slightly more complex than the others in that either **B0V** or **00V** is used, the choice being made so that the number of **B** pulses between consecutive **V** pulses is odd. These BNZS codes with $N = 3$, 6, or 8 involve bipolar violations and must therefore be carefully replaced by their equivalent zero strings at the receiver.

There are many other transmission codes, too numerous to list here. A list of codes and appropriate references can be found in Bylanski and Ingram.[4]

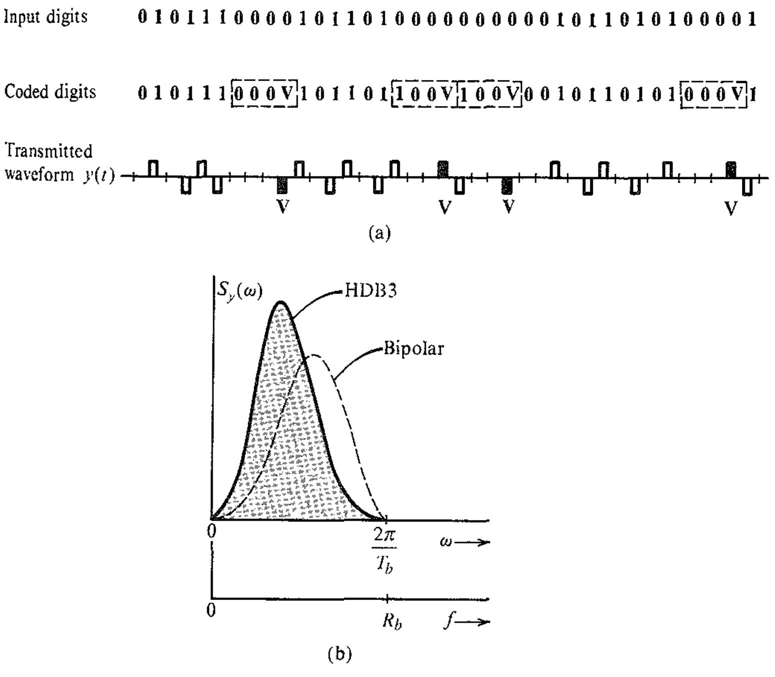


Figure 7.9 HDB3 signal and its PSD.

7.3 PULSE SHAPING

The PSD $S_y(\omega)$ of a digital signal $y(t)$ can be controlled by a choice of line code or by the pulse shape $P(\omega)$. In the last section , we discussed how the PSD is controlled by a line code. In this section, we examine how $S_y(\omega)$ is influenced by the pulse shape $p(t)$, and how to shape a pulse $p(t)$ in order to achieve a desired $S_y(\omega)$. The PSD $S_y(\omega)$ is strongly and directly influenced by the pulse shape $p(t)$ because $S_y(\omega)$ contains the term $|P(\omega)|^2$. Thus, compared to the nature of the line code, the pulse shape is a much more potent factor in terms of shaping the PSD $S_y(\omega)$.

In the last section, we used a simple half-width rectangular pulse $p(t)$ for the sake of illustration. Strictly speaking, in this case the bandwidth of $S_y(\omega)$ is infinite since $P(\omega)$ has infinite bandwidth. But we found that the essential bandwidth of $S_y(\omega)$ was finite. For example, most of the power of a bipolar signal is contained within the essential band of 0 to R_b Hz. Note, however, that the PSD is small but is still nonzero in the range of $f > R_b$ Hz. Therefore, when such a signal is transmitted over a channel of bandwidth R_b Hz, a significant portion of its spectrum is transmitted, but a small portion of the spectrum is suppressed. In Sec. 3.6, we saw how such a spectral distortion tends to spread the pulse (dispersion). Spreading of a pulse beyond its interval T_b will cause it to interfere with neighboring pulses. This is known as **intersymbol interference (ISI)**, which can cause errors in the correct detection of pulses.

To resolve the difficulty of ISI, let us review briefly our problem. We need to transmit a pulse every T_b interval, the kth pulse being $a_k p(t-kT_b)$. The channel has a finite bandwidth, and we are required to detect the pulse amplitude a_k correctly (that is, without ISI). In our discussion so far, we are considering time-limited pulses. Since such pulses cannot be band-limited, part of their spectra is suppressed by a band-limited channel. This causes pulse distortion (spreading out) and, consequently, ISI. We can try to resolve this difficulty by using pulses which are band-limited to begin with so that they can be transmitted intact over a band-limited channel. But band-limited pulses cannot be time-limited. Obviously, various pulses will overlap and cause ISI. Thus, whether we begin with time-limited pulses or band-limited pulses, it appears that ISI cannot be avoided. It is inherent in the finite transmission bandwidth. Fortunately, there is an escape from this blind alley. Pulse amplitudes can be detected correctly despite pulse spreading (or overlapping) if there is no ISI at the decision-making instants. This can be accomplished by a properly shaped band-limited pulse. To eliminate ISI, Nyquist proposed three different criteria for pulse shaping.[5] We shall consider only the first two criteria. The third is inferior to the first two[6], and, hence, will not be considered here.

7.3.1 Nyquist Criterion for Zero ISI

In the first method, Nyquist achieves zero ISI by choosing a pulse shape that has a nonzero amplitude at its center (say $t = 0$) and zero amplitudes at $t = \pm nT_b$ ($n = 1, 2, 3, \ldots$), where T_b is the separation between successive transmitted pulses (Fig. 7.10a),

$$p(t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm nT_b \end{cases} \quad \left(T_b = \frac{1}{R_b}\right) \tag{7.22}$$

A pulse satisfying this criterion causes zero ISI at all the remaining pulse centers, or signaling instants is shown in Fig. 7.10a, where we show several successive pulses (dotted) centered at

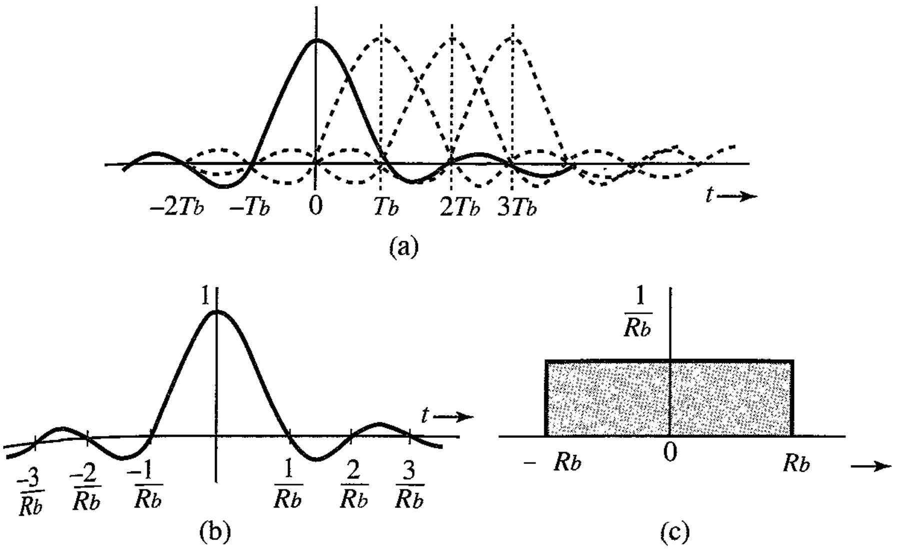


Figure 7.10 Minimum bandwidth pulse that satisfies the Nyquist criterion and its spectrum.

$t = 0,\ T_b,\ 2T_b,\ 3T_b,\ \ldots\ (T_b = 1/R_b)$. For the sake of convenience we have shown all pulses to be positive.* It is clear from this figure that the samples at $t = 0,\ T_b,\ 2T_b,\ 3T_b,\ \ldots$ consist of the amplitude of only one pulse (centered at the sampling instant) with no interference from the remaining pulses.

Now transmission of R_b bit/s requires a theoretical minimum bandwidth of $R_b/2$ Hz. It would be nice if a pulse satisfying Nyquist's criterion had this minimum bandwidth $R_b/2$ Hz. Can we find such a pulse $p(t)$? We have already solved this problem in Example 6.1 (with $B = R_b/2$), where we showed that there exists one (and only one) pulse that meets Nyquist's criterion (7.22) and has a bandwidth $R_b/2$ Hz. This pulse, $p(t) = \text{sinc}\,(\pi R_b t)$, (see Fig. 7.10b) has the property

$$\text{sinc}\,(\pi R_b t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm n T_b \end{cases} \quad \left(T_b = \frac{1}{R_b}\right) \tag{7.23a}$$

Moreover, the Fourier transform of this pulse is

$$P(\omega) = \frac{1}{R_b}\text{rect}\left(\frac{\omega}{2\pi R_b}\right) \tag{7.23b}$$

which has a bandwidth $R_b/2$ Hz;? as seen from Fig. 7.10c. Using this pulse, we can transmit at a rate of R_b pulses per second without ISI, over a bandwidth of $R_b/2$.

This scheme shows that we can attain the theoretical limit of performance by using a sinc pulse. Unfortunately, this pulse is impractical because it starts at $-\infty$. We will have to wait an infinite time to generate it. Any attempt to truncate it would increase its bandwidth beyond $R_b/2$ Hz. But even if this pulse were realizable, it has the undesirable feature that it decays too slowly at a rate $1/t$. This causes some serious practical problems. For instance, if the nominal data rate of R_b bit/s required for this scheme deviates a little, the pulse amplitudes will not vanish at the other pulse centers. Because the pulses decay only as $1/t$, the cumulative interference at any pulse center from all the remaining pulses is of the form $\sum(1/n)$. It is well known that an infinite series of this form does not converge and can add up to a very large value. A similar result occurs if everything is perfect at the transmitter, but the sampling rate at the receiver deviates from the rate of R_b Hz. Again, the same thing happens if the sampling instants deviate a little because of pulse time jitter, which is inevitable even in the most sophisticated systems. This scheme therefore fails unless everything is perfect, which is a practical impossibility. And all this is because sinc $(\pi R_b t)$ decays too slowly (as $1/t$). The solution is to find a pulse $p(t)$ that satisfies Eq. (7.22) but decays faster than $1/t$. Nyquist has shown that such a pulse requires a bandwidth of $kR_b/2$, with $1 \leq k \leq 2$.

---

* Actually, a pulse corresponding to **0** would be negative. But considering all positive pulses does not affect our reasoning. Showing negative pulses would make the figure needlessly confusing.

This can be proved as follows. Let $p(t) \Longleftrightarrow P(\omega)$, where the bandwidth of $P(\omega)$ is in the range $(R_b/2,\ R_b)$ (Fig. 7.11a). The desired pulse $p(t)$ satisfies Eq. (7.22). If we sample $p(t)$ every T_b seconds by multiplying $p(t)$ by $\delta_{T_b}(t)$ (an impulse train), then because of the property (7.22), all the samples, except the one at the origin, are zero. Thus, the sampled signal $\overline{p}(t)$ is

$$\overline{p}(t) = p(t)\delta_{T_b}(t) = \delta(t) \tag{7.24}$$

From Eq. (6.4) we know that the spectrum of a sampled signal $\overline{p}(t)$ is ($1/T_b$ times) the spectrum of $p(t)$ repeating periodically at intervals of the sampling frequency ω_b. Therefore, the Fourier transform of both sides of Eq. (7.24) yields

$$\frac{1}{T_b}\sum_{n=-\infty}^{\infty} P(\omega - n\omega_b) = 1 \qquad \omega_b = \frac{2\pi}{T_b} = 2\pi R_b \tag{7.25}$$

or

$$\sum_{n=-\infty}^{\infty} P(\omega - n\omega_b) = T_b \tag{7.26}$$

Thus, the sum of the spectra formed by repeating $P(\omega)$ every ω_b is a constant T_b, as shown in Fig. 7.11b.*

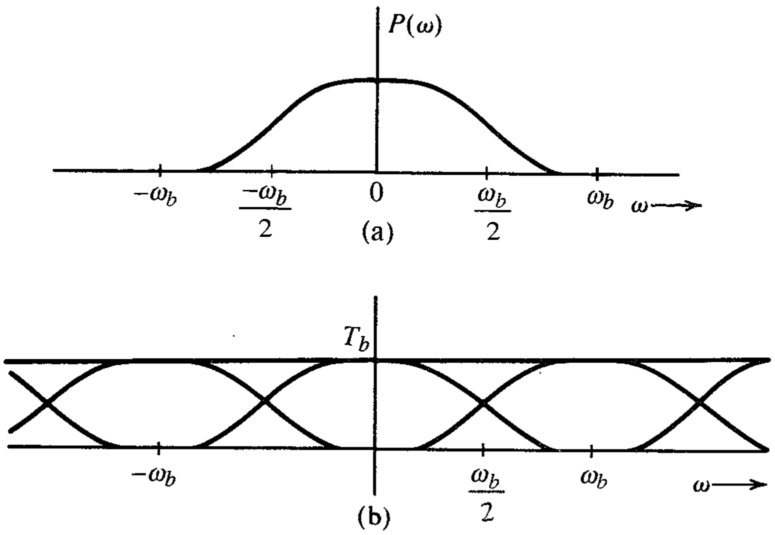


Figure 7.11 Derivation of the Nyquist criterion pulse.

---

* Observe that if $\omega_b > 2W$, where W is the bandwidth of $P(\omega)$, the repetitions of $P(\omega)$ are nonoverlapping, and the condition (7.26) cannot be satisfied. For $\omega_b = 2W$, the condition is satisfied only for the ideal low-pass $P(\omega)$ [$p(t) =$ sinc $(\pi R_b t)$], which is not acceptable. Hence, we must have $W > \omega_b/2$.

Consider the spectrum in Fig. 7.11b over the range $0 < \omega < \omega_b$. Over this range only the two terms $P(\omega)$ and $P(\omega - \omega_b)$ in the summation in Eq. (7.26) are involved. Hence,

$$P(\omega) + P(\omega - \omega_b) = T_b \qquad 0 < \omega < \omega_b$$

Letting $\omega = x + \omega_b/2$,

$$P\left(x + \frac{\omega_b}{2}\right) + P\left(x - \frac{\omega_b}{2}\right) = T_b \qquad |x| < \frac{\omega_b}{2} \tag{7.27}$$

Use of the result in Eq. (3.9) in Eq. (7.27) yields

$$P\left(\frac{\omega_b}{2} + x\right) + P^*\left(\frac{\omega_b}{2} - x\right) = T_b \qquad |x| < \frac{\omega_b}{2} \tag{7.28}$$

If we assume $P(\omega)$ of the form

$$P(\omega) = |P(\omega)|e^{-j\omega t_d}$$

then the term $e^{-j\omega t_d}$ represents pure time delay, and only $|P(\omega)|$ needs satisfy Eq. (7.28). Because $|P(\omega)|$ is real, Eq. (7.28) implies

$$\left|P\left(\frac{\omega_b}{2} + x\right)\right| + \left|P\left(\frac{\omega_b}{2} - x\right)\right| = T_b \qquad |x| < \frac{\omega_b}{2} \tag{7.29}$$

Hence, $|P(\omega)|$ should be of the form shown in Fig. 7.12. This curve has an odd symmetry about the set of axes intersecting at point x [the point on the $|P(\omega)|$ curve at $\omega = \omega_b/2$]. Note that this requires that $|P(\omega_b/2)| = 0.5|P(0)|$.

The bandwidth of $P(\omega)$ is $(\omega_b/2) + \omega_x$, where ω_x is the bandwidth in excess of the theoretical minimum bandwidth. Let r be the ratio of the excess bandwidth ω_x to the theoretical minimum bandwidth $\omega_b/2$;

$$\begin{aligned} r &= \frac{\text{excess bandwidth}}{\text{theoretical minimum bandwidth}} \\ &= \frac{\omega_x}{\omega_b/2} \\ &= \frac{2\omega_x}{\omega_b} \end{aligned} \tag{7.30}$$

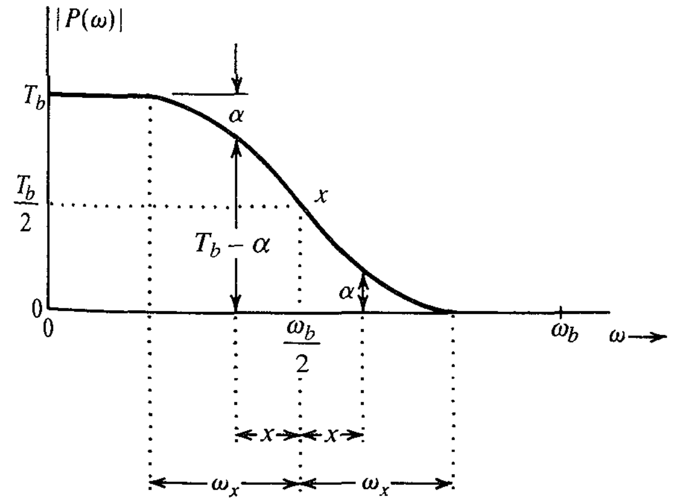


Figure 7.12 Vestigial spectrum.

Observe that because ω_x is at most equal to $\omega_b/2$,

$$0 \le r \le 1 \tag{7.31}$$

The theoretical minimum bandwidth is $R_b/2$ Hz, and the excess bandwidth is $f_x = rR_b/2$ Hz. Therefore, the bandwidth of $P(\omega)$ is

$$B_T = \frac{R_b}{2} + \frac{rR_b}{2} = \frac{(1+r)R_b}{2} \tag{7.32}$$

The constant r is called the **roll-off factor** and is also expressed in percent. For example, if $P(\omega)$ is a Nyquist criterion spectrum with a bandwidth that is 50% higher than the theoretical minimum, its roll-off factor $r = 0.5$, or 50%.

Although the phase of $P(\omega)$ must be linear up to the frequency where $|P(\omega)|$ goes to 0, for most practical applications it is sufficient to equalize the phase characteristics up to the 10- to 15-dB attenuation point [i.e., the point where $P(\omega)$ is 10 to 15 dB below its peak]. A filter having an amplitude response with the same characteristics is required in the vestigial sideband modulation discussed in Sec. 4.6. For this reason, we shall refer to the spectrum $P(\omega)$ in Eqs. (7.28) and (7.29) as a **vestigial spectrum**.

The pulse $p(t)$ in Eq. (7.22) has zero ISI at the centers of all other pulses transmitted at the rate of R_b pulses per second. A pulse $p(t)$ that causes zero ISI at the centers of all the remaining pulses (or signaling instants) is the Nyquist criterion pulse. We have shown that a pulse with a vestigial spectrum [Eq. (7.28) or Eq. (7.29)] satisfies the Nyquist criterion for zero ISI.

Because $0 \le r < 1$, the bandwidth of $P(\omega)$ is restricted to the range of $R_b/2$ to R_b Hz. The pulse $p(t)$ can be generated as a unit impulse response of a filter with transfer function $P(\omega)$. But because $P(\omega) = 0$ over a band, it violates the Paley-Wiener criterion and is therefore unrealizable. However, the vestigial roll-off characteristic is gradual, and it can be more closely approximated by a practical filter. One family of spectra that satisfies the Nyquist criterion is

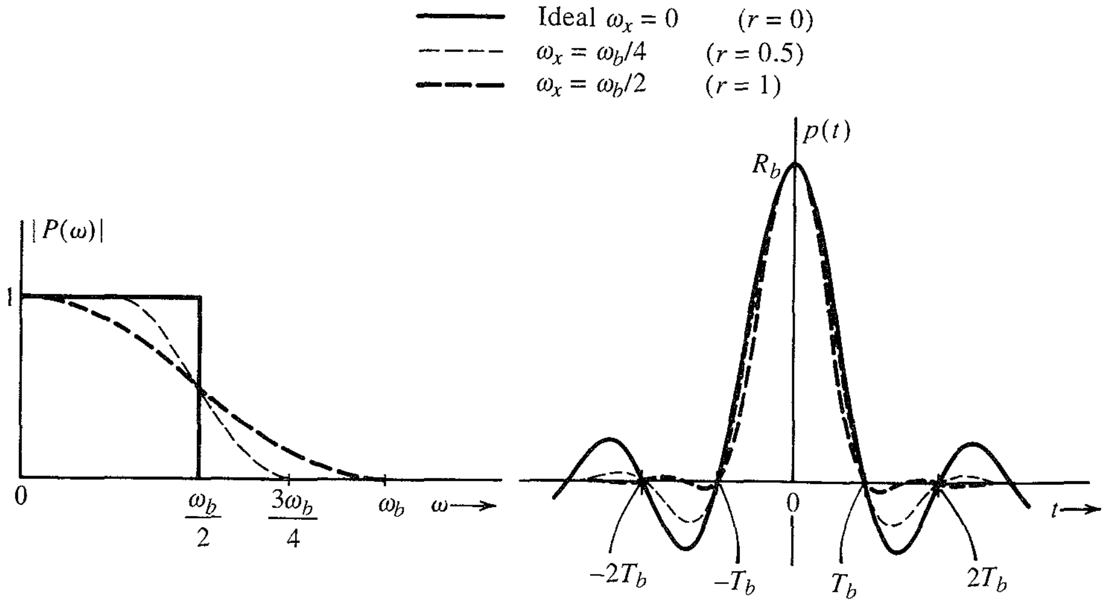
$$P(\omega) = \begin{cases} \dfrac{1}{2}\left\{1 - \sin\left(\dfrac{\pi[\omega - (\omega_b/2)]}{2\omega_x}\right)\right\} & \left|\omega - \dfrac{\omega_b}{2}\right| < \omega_x \\ 0 & |\omega| > \dfrac{\omega_b}{2} + \omega_x \\ 1 & |\omega| < \dfrac{\omega_b}{2} - \omega_x \end{cases} \tag{7.33}$$

Figure 7.13a shows three curves, corresponding to $\omega_x = 0\ (r = 0)$, $\omega_x = \omega_b/4\ (r = 0.5)$, and $\omega_x = \omega_b/2\ (r = 1)$. The respective impulse responses are shown in Fig. 7.13b. It can be seen that increasing ω_x (or r) improves $p(t)$; that is, more gradual cutoff reduces the oscillatory nature of $p(t)$ and causes it to decay more rapidly. For the case of the maximum value of $\omega_x = \omega_b/2\ (r = 1)$, Eq. (7.33) reduces to

$$P(\omega) = \frac{1}{2}\left(1 + \cos\frac{\omega}{2R_b}\right)\text{rect}\left(\frac{\omega}{4\pi R_b}\right) \tag{7.34a}$$

$$= \cos^2\left(\frac{\omega}{4R_b}\right)\text{rect}\left(\frac{\omega}{4\pi R_b}\right) \tag{7.34b}$$

This characteristic is known in the literature as the **raised-cosine** characteristic, because it represents a cosine raised by its peak amplitude. It is also known as the **full-cosine roll-**



(a) (b)

Figure 7.13 Pulses satisfying the Nyquist criterion.

off characteristic. The inverse Fourier transform of this spectrum is readily found as (see Prob. 7.3-7)

$$p(t) = R_b \frac{\cos \pi R_b t}{1 - 4R_b^2 t^2} \operatorname{sinc}(\pi R_b t) \tag{7.35}$$

This pulse is shown in Fig. 7.13b ($r = 1$). We can make several important observations about the raised-cosine pulse. First, the bandwidth of this pulse is R_b Hz and has the value R_b at $t = 0$ and zero not only at all the remaining signaling instants but also at points midway between all the signaling instants. Secondly, it decays rapidly, as $1/t^3$. As a result, the raised-cosine pulse is relatively insensitive to deviations of R_b, sampling rate, timing jitter, and so on. Furthermore, the pulse-generating filter with transfer function $P(\omega)$ [Eq. (7.34b)] is closely realizable. The phase characteristic that goes along with this filter is very nearly linear, so that no additional phase equalization is needed. Lastly, we shall see that the raised-cosine pulse can also be used as a duobinary pulse, which uses the principle of controlled ISI for correct detection.

It should be remembered that it is the pulses received at the detector input that should have the Nyquist form for zero ISI. In practice, because the channel is not ideal (distortionless), the transmitted pulses should be shaped so that after passing through the channel with transfer function $H_c(\omega)$, they will be received in the proper shape (such as raised-cosine pulses) at the receiver. Hence, the transmitted pulse $p_i(t)$ should satisfy

$$P_i(\omega) H_c(\omega) = P(\omega)$$

where $P(\omega)$ is the vestigial spectrum in Eq. (7.29).

EXAMPLE 7.1 Determine the pulse transmission rate in terms of the transmission bandwidth B_T and the roll-off factor r. Assume a scheme using the Nyquist criterion.

From Eq. (7.32),

$$R_b = \frac{2}{1+r} B_T$$

Because $0 \leq r \leq 1$, the pulse transmission rate varies from $2B_T$ to B_T, depending on the choice of r. A smaller r gives a higher signaling rate. But the pulse $p(t)$ decays slowly, creating the same problems as those discussed for the sinc pulse. For the raised-cosine pulse $r = 1$ and $R_b = B_T$, half the theoretical maximum rate. But the pulse decays faster as $1/t^3$ and is less vulnerable to ISI.

7.3.2 Signaling with Controlled ISI: Partial Response Signals

The Nyquist criterion pulse results in a bandwidth somewhat larger than the theoretical minimum. If we wish to reduce the pulse bandwidth further, we must somehow widen the pulse $p(t)$ (the wider the pulse, the narrower the bandwidth). Widening the pulse may result in interference (ISI) with the neighboring pulses. However, in the binary case with just two symbols, a known amount of ISI may be permissible because there are only a few possible interference patterns.

Consider a pulse specified by (see Fig. 7.14)

$$p(nT_b) = \begin{cases} 1 & n = 0,\ 1 \\ 0 & \text{for all other } n \end{cases} \tag{7.36}$$

We use polar signaling using this pulse. Thus, **1** is transmitted by $p(t)$ and **0** is transmitted by using the pulse $-p(t)$. The received signal is sampled at $t = nT_b$, and the pulse $p(t)$ has zero value at all n except $n = 0$ and 1, where its value is 1 (Fig. 7.14). Clearly, such a pulse causes zero ISI with all the pulses except the succeeding pulse. Therefore, we need worry about the ISI with the succeeding pulse only. Consider two such successive pulses located at 0 and T_b, respectively. If both pulses were positive, the sample value of the resulting signal at $t = T_b$ would be 2. If the both pulses were negative, the sample value would be -2. But if the two pulses were of opposite polarity, the sample value would be 0. This clearly allows us to make correct decisions at the sampling instants. The decision rule is as follows. If the sample value

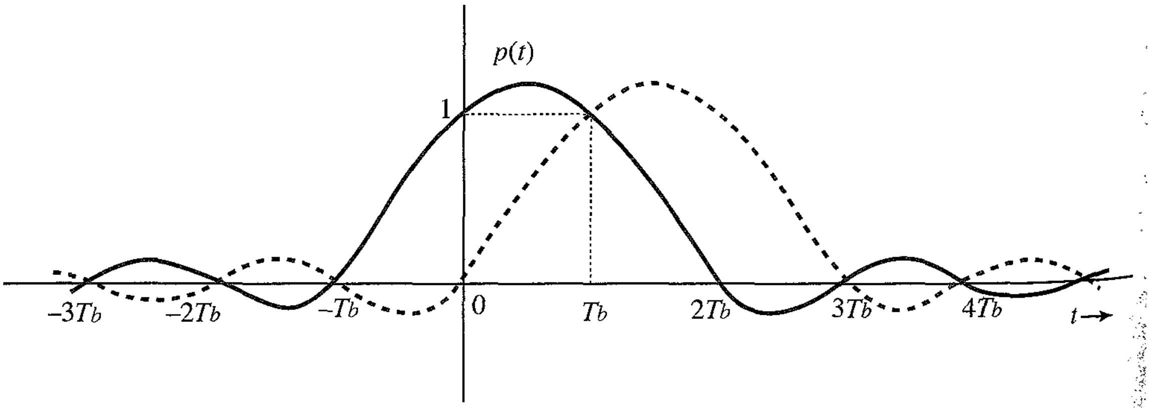


Figure 7.14 Communication using duobinary pulses.

is positive, the present bit is **1** and the previous bit is also **1**. If the sample value is negative, the present bit is **0** and the previous bit is also **0**. If the sample value is zero, the present bit is the complement of the previous bit. The knowledge of the previous bit then allows the determination of the present bit.

Figure 7.15 shows a transmitted bit sequence, the sample values of the received signal $x(t)$ (assuming no errors caused by channel noise), and the detector decision. This example also indicates the error-detecting property of this scheme. Examination of samples of the waveform $x(t)$ in Fig. 7.15 shows that there are always an even number of zero-valued samples between two full-valued samples of the same polarity and an odd number of zero-valued samples between two full-valued samples of opposite polarity. Thus, the second sample value of $x(t)$ is 2, and the next full-valued sample (the fifth sample) is 2. Between these full-valued samples of the same polarity, there are an even number (i.e., 2) of zero-valued samples. If one of the sample values is detected wrong, this rule is violated, and the error is detected.

The pulse $p(t)$ goes to zero at $t = -T_b$ and $2T_b$, resulting in a pulse width (of the primary lobe) 50% higher than that of the Nyquist criterion pulse. This results in a reduction of its bandwidth. This is the second method proposed by Nyquist. This scheme of controlled ISI is also known as **correlative** or **partial-response** scheme. A pulse satisfying the criterion in Eq. (7.36) is called **duobinary pulse.**

Example of a Duobinary Pulse

If we restrict the pulse bandwidth to $R_b/2$, then following the procedure of Example 6.1, we can show that (see Prob. 7.3-9) only the following pulse $p(t)$ meets the requirement in Eq. (7.36) for the duobinary pulse,

$$p(t) = \frac{\sin(\pi R_b t)}{\pi R_b t(1 - R_b t)} \tag{7.37}$$

The Fourier transform $P(\omega)$ of the pulse $p(t)$ is given by (see Prob. 7.3-8)

$$P(\omega) = \frac{2}{R_b} \cos\left(\frac{\omega}{2R_b}\right) \operatorname{rect}\left(\frac{\omega}{2\pi R_b}\right) e^{-j\frac{\omega}{2R_b}} \tag{7.38}$$

The pulse $p(t)$ and its amplitude spectrum $|P(\omega)|$ are shown in Fig. 7.16.* This pulse transmits binary data at a rate of R_b bit/s, and has the theoretical minimum bandwidth $R_b/2$ Hz. Equation (7.37) shows that this pulse decays rapidly with time as $1/t^2$. This pulse is not realizable because $p(t)$ is noncausal and has infinite duration [since $P(\omega)$ is band-limited]. However, as it decays rapidly (as $1/t^2$), it can be closely approximated.

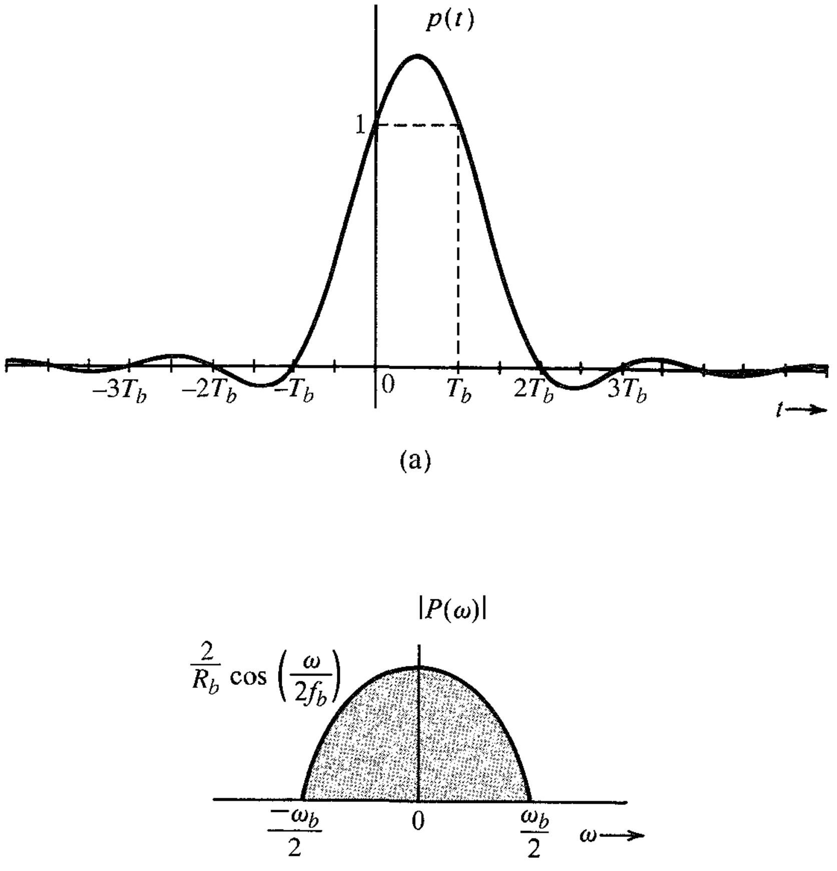
A careful examination of the raised-cosine pulse (Fig. 7.13 for $\omega_x = \omega_b/2$) shows that it satisfies the criterion in Eq. (7.36) also. Thus, data transmitted by a raised-cosine pulse can

Transmitted sequence	**1**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**1**
Samples of $x(t)$	1	2	0	0	2	0	−2	−2	0	0	0	2	2
Detected sequence	**1**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**1**

Figure 7.15 Transmitted bits and the received samples in controlled ISI signaling.

---

* The phase spectrum is linear, $\theta_p(\omega) = -\omega/2R_b$.



(a)

(b)

Figure 7.16 Minimum-bandwidth pulse that satisfies the duobinary pulse criterion and its spectrum.

be detected using the Nyquist criterion [sampling $x(t)$ at the signaling instants] or using the controlled ISI (partial response) method. But it uses twice the bandwidth required for $p(t)$ in Eq. (7.37).

It may come as a surprise that we are able to achieve the theoretical rate using the duobinary pulse. In fact, it is an illusion. The theoretical rate of transmission is two pieces of independent information per second per hertz bandwidth. We have achieved this rate for binary information. Here is the catch! A binary information does not qualify as an independent piece of information because it cannot take on an arbitrary value. It must be selected from a finite set. The duobinary pulse would fail if the pulses were truly independent pieces of information, that is, if the pulses were to have arbitrary amplitudes. The scheme works only because the binary pulses take on finite known values, and, hence, there is only a finite (known) number of interference patterns between pulses, which permits correct determination of pulse amplitudes despite interference.

Use of Differential Coding

For the controlled ISI method, Fig. 7.15 shows that a zero-valued sample implies transition, that is, if a digit is detected as **1**, the previous digit is **0**, or vice versa. This means the digit interpretation is based on the previous digit. If a digit were detected wrong, the error would tend to propagate. Use of the so-called **differential coding** eliminates this problem.

In differential coding a **1** is transmitted by a pulse identical to that used for the previous bit and a **0** is transmitted by a pulse negative of that used for the previous bit. This is shown in Fig. 7.17 using a half-width rectangular pulse. Observe that changing the polarity of this signal will not affect its interpretation. Such a coding is useful in systems that have no sense of absolute polarity.

If we use a duobinary pulse $p(t)$ instead of the rectangular pulse in this differential coding, something interesting happens. Suppose the kth data bit is **1**. Because of differential coding, this bit is transmitted by a pulse identical to the previous pulse (two pulses of the same polarity). Hence, the kth sample of the received signal is either 2 or -2. On the other hand, if the kth bit is **0**, the present pulse is the negative of the previous pulse (transition), and the kth sample of the received signal is 0. Thus, if we use differential coding in conjunction with the duobinary pulse, our detection procedure is simplified. If the sample value is 0, the incoming data bit is **0** and if the sample value is ± 2, the incoming data bit is **1**. This scheme not only simplifies the decision rule but also makes the decision independent of the previous digit and eliminates error propagation.

Pulse Generation

A pulse $p(t)$ satisfying a Nyquist criterion can be generated as the unit impulse response of a filter with transfer function $P(\omega)$. This may not always be easy. A better method is to generate the waveform directly, using a transversal filter (tapped delay line) encountered in Sec. 6.3. The pulse $p(t)$ to be generated is sampled with a sufficiently small sampling interval T_s (Fig. 7.18a), and the filter tap gains are set in proportion to these sample values in sequence, as shown in Fig. 7.18b. When a narrow rectangular pulse of width T_s, the sampling interval, is applied at the input of the transversal filter, the output will be a staircase approximation of $p(t)$. This output, when passed through a low-pass filter, is smoothed out. The approximation can be improved by reducing the pulse sampling interval T_s.

It should be stressed once again that the pulses arriving at the detector input of the receiver need to have the Nyquist or duobinary form. Hence, the transmitted pulses should be so shaped that after passing through the channel, they are received in the form of the desired (Nyquist or duobinary) form. In practice, however, pulses need not be shaped rigidly at the transmitter. The final shaping can be carried out by an equalizer at the receiver, as will be discussed in Sec. 7.5.

7.4 SCRAMBLING

In general, a scrambler tends to make the data more random by removing long strings of **1**'s or **0**'s. Scrambling can be helpful in timing extraction by removing long strings of **0**'s in binary

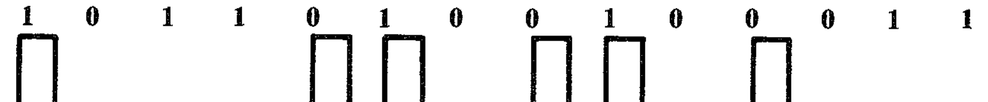


Figure 7.17 Differential code.

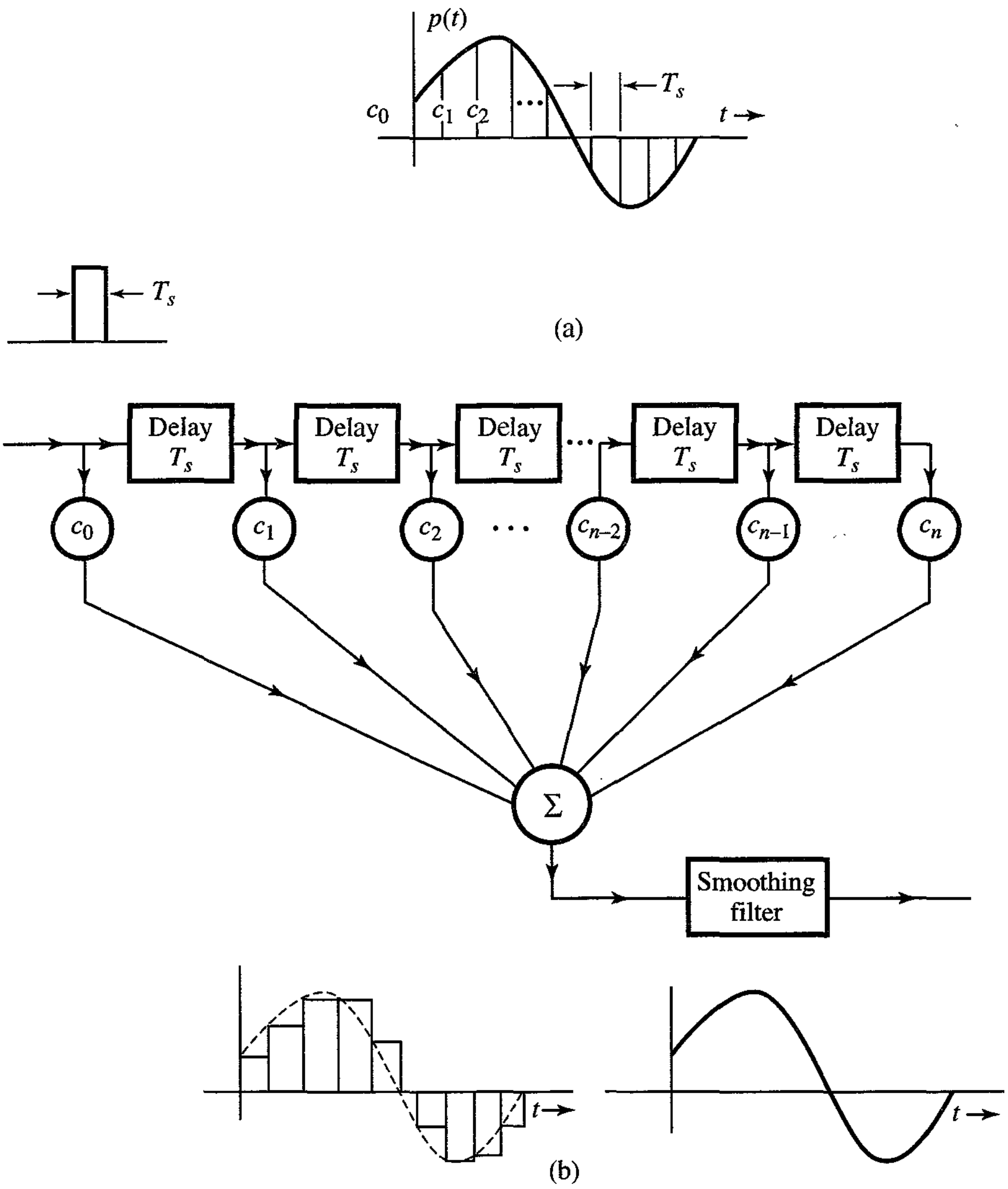


Figure 7.18 Pulse generation by transversal filter.

data. Scramblers, however, are primarily used for preventing unauthorized access to the data, and are optimized for that purpose. Such optimization may actually result in the generation of a long string of zeros in the data. The digital network must be able to cope with these long zero strings using zero suppression techniques as discussed in Sec. 7.2.

Figure 7.19 shows a typical scrambler and descrambler. The scrambler consists of a feedback shift register, and the matching descrambler has a feedforward shift register, as illustrated in Fig. 7.19. Each stage in the shift register delays a bit by one unit. To analyze the scrambler and the matched descrambler, consider the output sequence T of the scrambler (Fig. 7.19a). If S is the input sequence to the scrambler, then

$$S \oplus D^3T \oplus D^5T = T \tag{7.39}$$

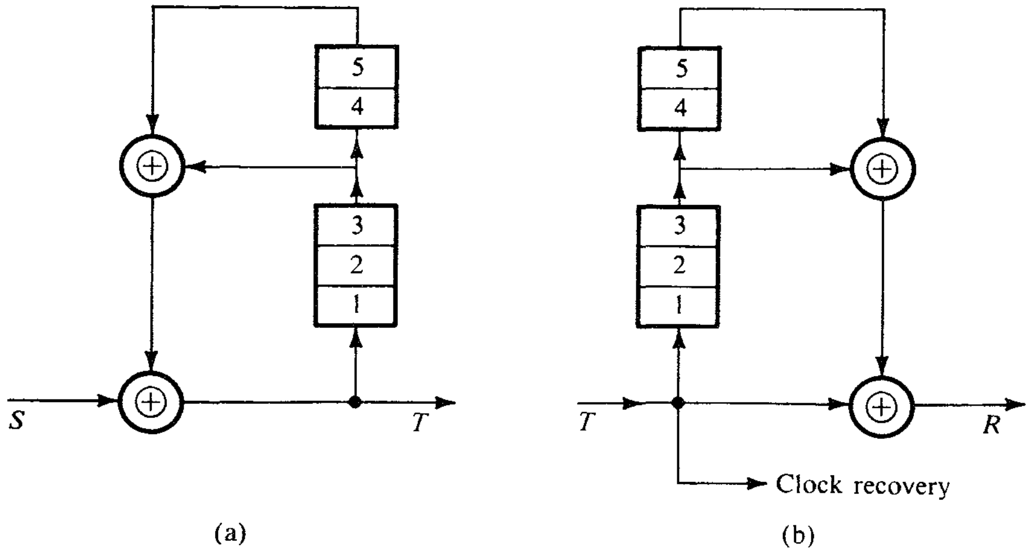


Figure 7.19 Scrambler and descrambler.

where D represents the delay operator; that is, $D^n T$ is the sequence T delayed by n units. The symbol $\oplus$ indicates modulo 2 sum. Now, recall that the modulo 2 sum of any sequence with itself gives a sequence of all **0**'s. Adding $(D^3 \oplus D^5)T$ to both sides of Eq. (7.39), we get

$$\begin{aligned} S &= T \oplus (D^3 \oplus D^5)T \\ &= [1 \oplus (D^3 \oplus D^5)]T \\ &= (1 \oplus F)T \qquad \text{where } F = D^3 \oplus D^5 \end{aligned} \tag{7.40}$$

To design the descrambler at the receiver, we start with T, the sequence received at the descrambler. From Eq. (7.40), it follows that

$$S = T \oplus FT = T \oplus (D^3 \oplus D^5)T$$

This equation, where we regenerate the input sequence S from the received sequence T, is readily implemented by the descrambler shown in Fig. 7.19b.

Note that a single detection error in the received sequence T will affect three output bits in R. Hence, scrambling has the disadvantage of causing multiple errors for a single received bit error.

EXAMPLE 7.2 The data stream **101010100000111** is fed to the scrambler in Fig. 7.19a. Find the scrambler output T, assuming the initial content of the registers to be zero.

From Fig. 7.19a we observe that initially $T = S$, and the sequence S enters the register and is returned as $(D^3 \oplus D^5)S = FS$ through the feedback path. This new sequence FS again enters the register and is returned as F^2S, and so on. Hence,

$$\begin{aligned} T &= S \oplus FS \oplus F^2S \oplus F^3S \oplus \cdots \\ &= (1 \oplus F \oplus F^2 \oplus F^3 \oplus \cdots)S \end{aligned} \tag{7.41}$$

Recognizing that

$$F = D^3 \oplus D^5$$

we have

$$F^2 = (D^3 \oplus D^5)(D^3 \oplus D^5) = D^6 \oplus D^{10} \oplus D^8 \oplus D^8$$

Because modulo-2 addition of any sequence with itself is zero, $D^8 \oplus D^8 = 0$, and

$$F^2 = D^6 \oplus D^{10}$$

Similarly,

$$F^3 = (D^6 \oplus D^{10})(D^3 \oplus D^5) = D^9 \oplus D^{11} \oplus D^{13} \oplus D^{15}$$

and so on. Hence [see Eq. (7.41)],

$$T = (1 \oplus D^3 \oplus D^5 \oplus D^6 \oplus D^9 \oplus D^{10} \oplus D^{11} \oplus D^{12} \oplus D^{13} \oplus D^{15} \oplus \cdots)S$$

Because $D^n S$ is simply the sequence S delayed by n bits, various terms in the preceding equation correspond to the following sequences:

$$\begin{aligned}
S &= \mathbf{101010100000111} \\
D^3 S &= \mathbf{000101010100000111} \\
D^5 S &= \mathbf{00000101010100000111} \\
D^6 S &= \mathbf{000000101010100000111} \\
D^9 S &= \mathbf{000000000101010100000111} \\
D^{10} S &= \mathbf{0000000000101010100000111} \\
D^{11} S &= \mathbf{00000000000101010100000111} \\
D^{12} S &= \mathbf{000000000000101010100000111} \\
D^{13} S &= \mathbf{0000000000000101010100000111} \\
\underline{D^{15} S} &= \underline{\mathbf{000000000000000}}\mathbf{101010100000111} \\
T &= 101110001101001
\end{aligned}$$

Note that the input sequence contains the periodic sequence **10101010** . . . , as well as a long string of **0**'s. The scrambler output effectively removes the periodic component as well as the long string of **0**'s. The input sequence has 15 digits. The scrambler output up to the 15th digit only is shown, because all the output digits beyond 15 depend on the input digits beyond 15, which are not given.

We can verify that the descrambler output is indeed S when this sequence T is applied at its input (see Prob. 7.4-1).

7.5 REGENERATIVE REPEATER

Basically, a regenerative repeater performs three functions: (1) reshaping incoming pulses by means of an equalizer, (2) the extraction of timing information required to sample incoming pulses at optimum instants, and (3) decision making based on the pulse samples. The schematic of a repeater is shown in Fig. 7.20. A complete repeater also includes provision for the separation of dc power from ac signals. This is normally accomplished by transformer coupling the signals and bypassing the dc around the transformers to the power supply circuitry.*

Preamplifier and Equalizer

A pulse train is attenuated and distorted by the transmission medium. The distortion is in the form of dispersion, which is caused by an attenuation of high-frequency components of the pulse train. Theoretically, an equalizer should have a frequency characteristic that is the inverse of that of the transmission medium. This will restore higher frequency components and eliminate pulse dispersion. Unfortunately, this also increases the received channel noise by boosting its high-frequency components. For digital signals, however, complete equalization is really not necessary, because a detector has to make relatively simple decisions—such as whether the pulse is positive or negative (or whether the pulse is present or absent). Therefore, considerable pulse dispersion can be tolerated. Pulse dispersion results in ISI and the consequent increase in detection error probability. Noise increase resulting from the equalizer (which boosts the high frequencies) also increases the detection error probability. For this reason, the design of an optimum equalizer involves an inevitable compromise between reducing ISI and reducing the channel noise. A judicious choice of the equalization characteristics is a central feature of all digital communication systems.

Zero-Forcing Equalizer: It is really not necessary to eliminate or minimize ISI (interference) with neighboring pulses for all t. All that is needed is to eliminate or minimize interference with neighboring pulses at their respective sampling instants only, because the decision is based only on sample values. This can be accomplished by the transversal-filter equalizer encountered earlier, which forces the equalizer output pulse to have zero values at

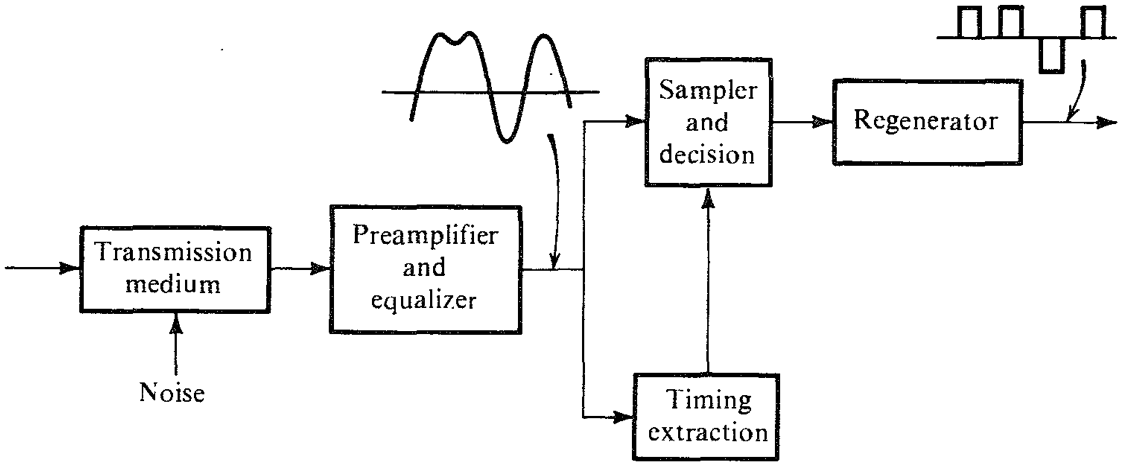


Figure 7.20 Regenerative repeater.

---

* The repeater usually includes circuitry to protect the electronics of the regenerator from power and lightning induced high-voltage transients. Special transformer windings may be provided to couple fault locate signals into a cable pair dedicated to the purpose.

the sampling (decision-making) instants. In other words, the equalizer output pulses should satisfy the Nyquist criterion or the controlled ISI criterion. The time delay T between successive taps is chosen to be T_b, the interval between pulses.

To begin with, set the tap gains $c_0 = 1$ and $c_k = 0$ for all other values of k in the transversal filter in Fig. 7.21a. Thus the output of the filter will be the same as the input delayed by NT_b. For a single pulse $p_r(t)$ (Fig. 7.21b) at the input of the transversal filter with the above tap setting, the filter output $p_o(t)$ will be exactly $p_r(t - NT_b)$, that is, $p_r(t)$ delayed by NT_b. This delay is not relevant to our discussion. Hence, for convenience, we shall ignore this delay. This means that $p_r(t)$ in Fig. 7.21b also represents the filter output $p_o(t)$ for this tap setting ($c_0 = 1$ and $c_k = 0$, $k \neq 0$). We require that the output pulse $p_o(t)$ satisfy the Nyquist criterion or the controlled ISI criterion, as the case may be. For the Nyquist criterion, the output pulse $p_o(t)$ must have zero values at all the multiples of T_b. From Fig. 7.21b, we see

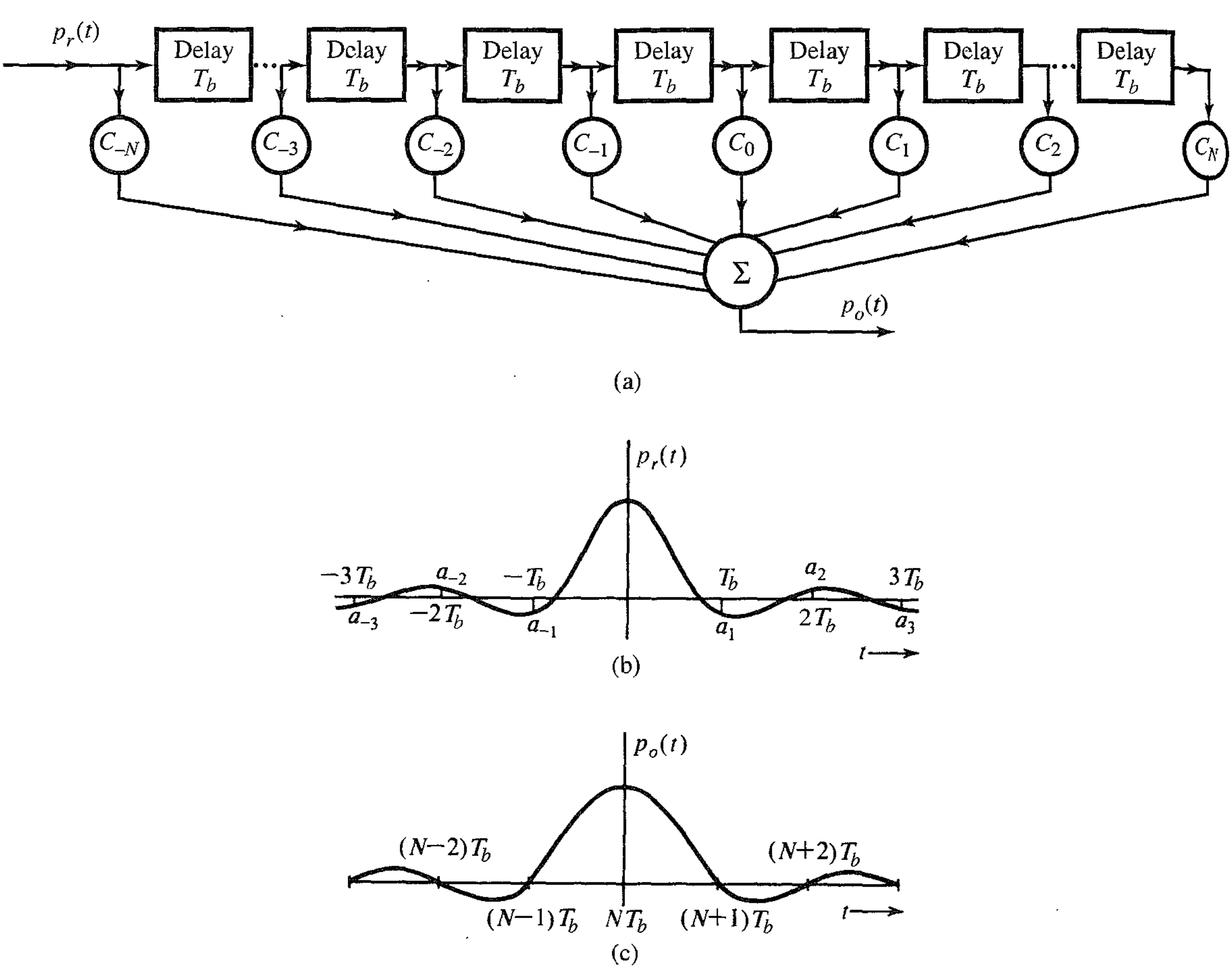

$p_r(t)$
Delay T_b
Delay T_b
Delay T_b
Delay T_b
Delay T_b
Delay T_b
Delay T_b
C_{-N}
C_{-3}
C_{-2}
C_{-1}
C_0
C_1
C_2
C_N
Σ
$p_o(t)$
(a)
$p_r(t)$
$-3T_b$
a_{-2}
$-T_b$
T_b
a_2
$3T_b$
a_{-3}
$-2T_b$
a_{-1}
a_1
$2T_b$
a_3
t
(b)
$p_o(t)$
$(N-2)T_b$
$(N+2)T_b$
$(N-1)T_b$
NT_b
$(N+1)T_b$
t
(c)

Figure 7.21 Zero-forcing equalizer analysis.

that the pulse amplitudes a_1, a_{-1}, and a_2 at T_b, $-T_b$, and $2T_b$, respectively, are not negligible. By adjusting the tap gains (c_k's), we generate additional shifted pulses of proper amplitudes that will force the resulting output pulse to have desired values at $t = 0,\ T_b,\ 2T_b,\ \ldots$. The output $p_o(t)$ (Fig 7.21c) is the sum of pulses of the form $c_k p_r(t - kT_b)$ (ignoring the delay of NT_b). Thus,

$$p_o(t) = \sum_{n=-N}^{N} c_n p_r(t - nT_b) \tag{7.42}$$

The samples of $p_o(t)$ at $t = kT_b$ are

$$p_o(kT_b) = \sum_{n=-N}^{N} c_n p_r[(k-n)T_b] \qquad k = 0,\ \pm1,\ \pm2,\ \pm3, \ldots \tag{7.43}$$

Using a more convenient notation $p_r[k]$ to denote $p_r(kT_b)$ and $p_o[k]$ to denote $p_o(kT_b)$, Eq. (7.43a) can be expressed as

$$p_o[k] = \sum_{n=-N}^{N} c_n p_r[k-n] \qquad k = 0,\ \pm1,\ \pm2,\ \pm3,\ \ldots \tag{7.43b}$$

The Nyquist criterion requires the samples $p_o[k] = 0$ for $k \neq 0$, and $p_o[k] = 1$ for $k = 0$. Substituting these values into Eq. (7.43b), we obtain a set of infinite simultaneous equations in terms of $2N + 1$ variables. Clearly, it is not possible to solve this set of equations. However, if we specify the values of $p_o[k]$ only at $2N + 1$ points as

$$p_o[k] = \begin{cases} 1 & k = 0 \\ 0 & k = \pm1,\ \pm2,\ \ldots,\ \pm N \end{cases} \tag{7.44}$$

then a unique solution exists. This assures that a pulse will have zero interference at the sampling instants of N preceding and N succeeding pulses. Because the pulse amplitude decays rapidly, interference beyond the Nth pulse is not significant for $N > 2$, in general. Substitution of the condition (7.44) into Eq. (7.43b) yields a set of $2N + 1$ simultaneous equations in $2N + 1$ variables:

$$\begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 0 \\ 1 \\ 0 \\ \cdots \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} p_r[0] & p_r[-1] & \cdots & p_r[-2N] \\ p_r[1] & p_r[0] & \cdots & p_r[-2N+1] \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ p_r[N-1] & p_r[N-2] & \cdots & p_r[-N-1] \\ p_r[N] & p_r[N-1] & \cdots & p_r[-N] \\ p_r[N+1] & p_r[N] & \cdots & p_r[-N+1] \\ \cdots & \cdots & \cdots & \cdots \\ p_r[2N-1] & p_r[2N-2] & \cdots & p_r[1] \\ p_r[2N] & p_r[2N-1] & \cdots & p_r[0] \end{bmatrix} \begin{bmatrix} c_{-N} \\ c_{-N+1} \\ \cdots \\ \\ c_{-1} \\ c_0 \\ c_1 \\ \cdots \\ c_{N-1} \\ c_N \end{bmatrix} \tag{7.45}$$

The tap-gain c_k's can be obtained by solving this set of equations.

EXAMPLE 7.3 For the received pulse $p_r(t)$ in Fig. 7.21b, let

$$a_0 = p_r[0] = 1$$
$$a_1 = p_r[1] = -0.3, \quad a_2 = p_r[2] = 0.1$$
$$a_{-1} = p_r[-1] = -0.2, \quad a_{-2} = p_r[-2] = 0.05$$

Design a three-tap ($N = 1$) equalizer.

Substituting the preceding values into Eq. (7.45), we obtain

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & -0.2 & 0.05 \\ -0.3 & 1 & -0.2 \\ 0.1 & -0.3 & 1 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix}$$

Solution of this set yields $c_{-1} = 0.210$, $c_0 = 1.13$, and $c_1 = 0.318$. This tap setting assures us that $p_r[0] = 1$ and $p_r[-1] = p_r[1] = 0$. The output $p_o(t)$ is sketched in Fig. 7.21c.

Least Mean Squared Error Equalizer: Another approach to equalization, the **least mean-squared error (LMSE)** method, does not try to force the pulse samples to zero at $2N$ points. Instead the mean of the squared errors over a set of output samples is minimized. This method also involves solving simultaneous equations.

Automatic and Adaptive Equalization: The setting of the tap gains of an equalizer can be done automatically by using a special sequence of pulses prior to the data transmission and by using an iterative technique to obtain optimum tap gains. In adaptive equalizers, the tap gains are adjusted continuously during transmission.[7,8]

Eye Diagram

The ISI and other signal degradations can be studied conveniently on an oscilloscope through what is known as the **eye diagram**. A random binary pulse sequence is sent over the channel. The channel output is applied to the vertical input of an oscilloscope. The time base of the scope is triggered at the same rate as that of the incoming pulses, and it yields a sweep lasting exactly T_b, the interval of one pulse. The oscilloscope shows the superposition of several traces, which is nothing but the input signal (vertical input) cut up every T_b and then superimposed (Fig. 7.22). The oscilloscope pattern thus formed looks like a human eye and, hence, the name eye diagram.

As an example, consider the transmission of a binary signal by polar rectangular pulses. If the channel is ideal with infinite bandwidth, pulses will be received without distortion. When this signal is cut up, every pulse interval or each piece will be either a positive or a negative rectangular pulse. When those are superimposed, the resulting eye diagram will be as shown in Fig. 7.22a. If the channel is not distortionless or has finite bandwidth, or both, received pulses will no longer be rectangular but will be rounded and spread out. If the equalizer is adjusted properly to eliminate ISI at the pulse sampling instants, the resulting eye diagram will be rounded (Fig. 7.22b) but will still have full opening at the midpoint of the eye. This is because the midpoint of the eye represents the sampling instant of each pulse, where the pulse

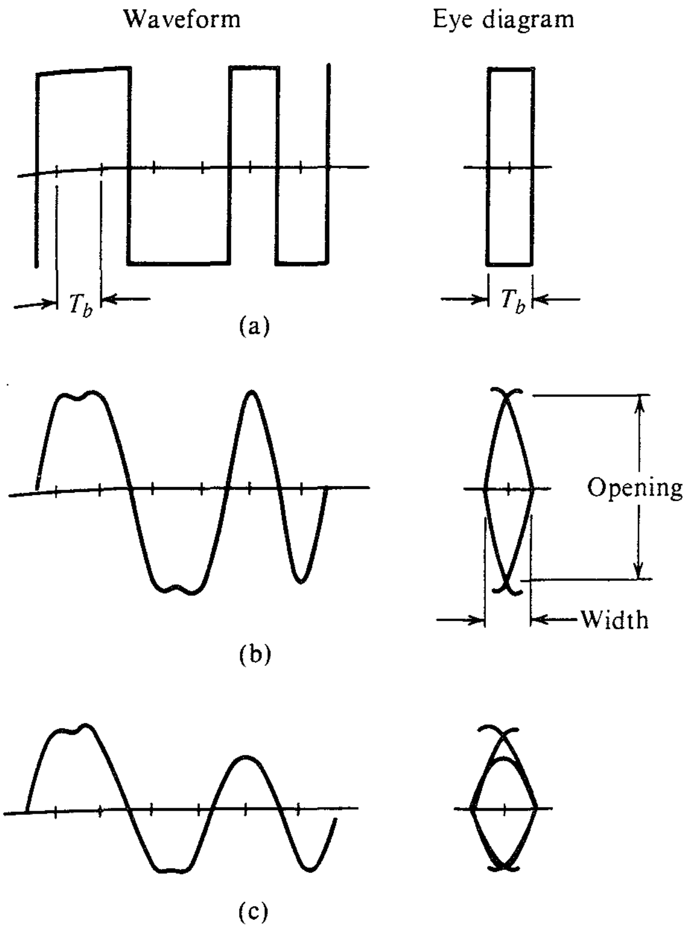


Figure 7.22 Eye diagram.

amplitude is maximum without interference from any other pulse (because of zero ISI). If ISI is not zero, pulse values at their respective sampling instants will deviate from the full-scale values by varying amounts in each trace, causing a blur and thus closing the eye partially at the midpoint, as illustrated in Fig. 7.22c.

In the presence of channel noise, the eye will tend to close in all cases. Smaller noise will cause proportionately less closing. The decision threshold as to which symbol (**1** or **0**) is transmitted is the midpoint of the eye.* Observe that for zero ISI, the system can tolerate noise of up to one-half the vertical opening of the eye. Because the ISI reduces the eye opening, it clearly reduces noise tolerance. The eye diagram is also used to determine optimum tap settings of the equalizer. Taps are adjusted to obtain the maximum vertical and horizontal eye openings.

The eye diagram is useful in deciding the optimum sampling or decision-making instant (the instant when the eye opening is maximum), as well as the amount of noise that can be tolerated. The width of the eye indicates the time interval over which the decision can be made, and it is desirable to have an eye with the maximum horizontal opening. If the decision-making instant deviates from the instant when the eye has a maximum vertical opening, the margin to noise tolerance is reduced. This causes higher error probability in pulse detection. Because in

---

[*] This is true for a two-level decision, e.g., when $p(t)$ and $-p(t)$ are used for **1** and **0**, respectively. For a three-level decision (bipolar signaling, for example), there will be two thresholds.

any system the sampling instants deviate from the ideal (because of the presence of jitter), the eye diagram allows one to study the effects of jitter.

Timing Extraction

The received digital signal needs to be sampled at precise instants. This requires a clock signal at the receiver in synchronism with the clock signal at the transmitter (**symbol** or **bit synchronization**). Three general methods of synchronization exist:

1. Derivation from a primary or a secondary standard (e.g., transmitter and receiver slaved to a master timing source).
2. Transmitting a separate synchronizing signal (pilot clock).
3. Self-synchronization, where the timing information is extracted from the received signal itself.

The first method is suitable for large volumes of data and high-speed communication systems because of its high cost. In the second method, part of the channel capacity is used to transmit timing information and is suitable when the available capacity is large compared to the data rate. The third method is a very efficient method of timing extraction or clock recovery because the timing is derived from the digital signal itself. An example of the self-synchronization method will be discussed here.

We have already shown that a digital signal, such as an on-off signal (Fig. 7.2), contains a discrete component of the clock frequency itself (Fig. 7.7). Hence, when the on-off binary signal is applied to a resonant circuit tuned to the clock frequency, the output signal is the desired clock signal.

Not all the binary signals contain a discrete component of the clock frequency. For example, a bipolar signal has no discrete component of any frequency [see Eqs. (7.20) or Fig. 7.8]. In such cases, it may be possible to extract timing by using a nonlinear operation. In the bipolar case, for instance, a simple rectification converts a bipolar signal to an on-off signal, which can readily be used to extract timing.

Small random deviations of the incoming pulses from their ideal locations (known as **timing jitter**) are always present, even in the most sophisticated systems. Although the source emits pulses at the right instants, subsequent operations during transmission (e.g., at repeaters) tend to deviate pulses from these original positions. The Q of the tuned circuit used for timing extraction must be large enough to provide an adequate suppression of timing jitter, yet small enough to meet the stability requirements. During the intervals where there are no pulses in the input, the oscillation continues because of the flywheel effect of the high-Q circuit. But still the oscillator output is sensitive to the pulse pattern. For example, during a long string of **1**'s the output amplitude will increase, whereas during a long string of **0**'s it will decrease. This introduces additional jitter in the timing signal extracted.

The complete timing extractor and time-pulse generator for a polar case in shown in Fig. 7.23. The sinusoidal output of the oscillator (timing extractor) is passed through a phase shifter that adjusts the phase of the timing signal so that the timing pulses occur at the maximum eye opening. This method is used to recover the clock at each of the regenerators in a PCM system. The jitter introduced by successive regenerators adds up, and after a certain number of regenerators it is necessary to use a regenerator with a more sophisticated clock recovery system such as a PLL.

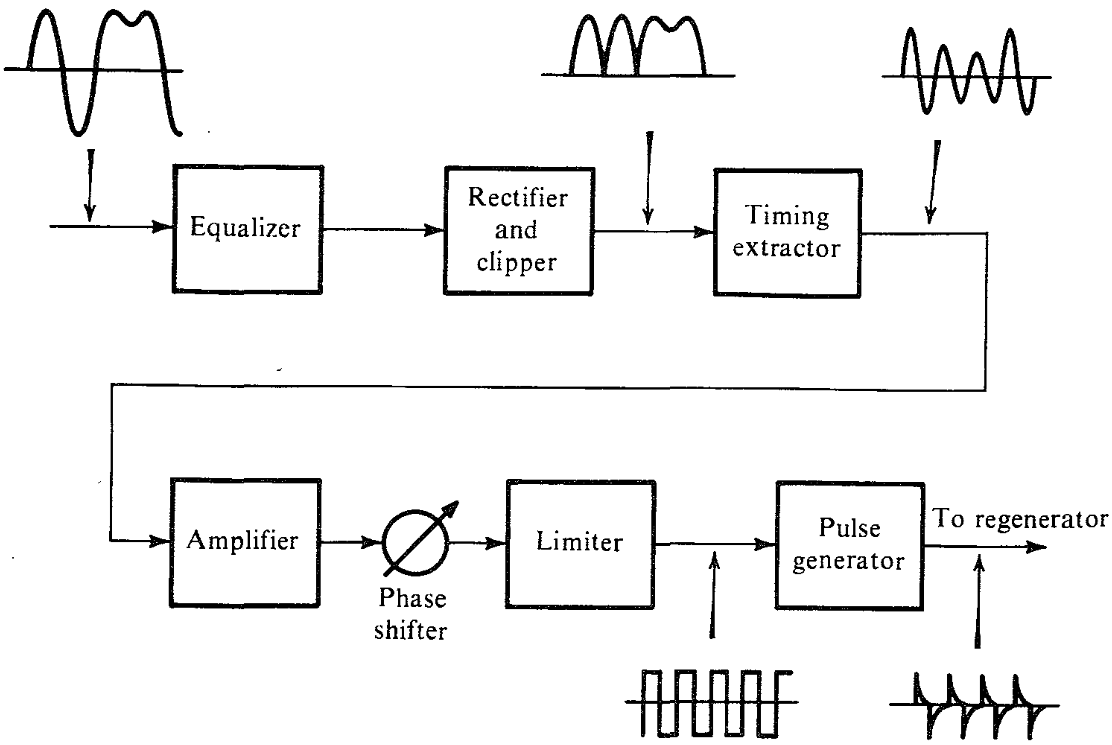


Figure 7.23 Timing extraction.

Timing Jitter: Variations of the pulse positions or sampling instants cause timing jitter. This results from several causes, some of which are dependent on the pulse pattern being transmitted whereas others are not. The former are cumulative along the chain of regenerative repeaters because all the repeaters are affected in the same way, whereas the other forms of jitter are random from regenerator to regenerator and therefore tend to partially cancel out their mutual effects over a long-haul link. Random forms of jitter are caused by noise, interference, and mistuning of the clock circuits. The pattern-dependent jitter results from clock mistuning, amplitude-to-phase conversion in the clock circuit, and ISI, which alters the position of the peaks of the input signal according to the pattern. The rms value of the jitter over a long chain of N repeaters can be shown to increase as $\sqrt{N}$.

Jitter accumulation over a digital link may be reduced by buffering the link with an elastic store and clocking out the digit stream under the control of a highly stable PLL. Jitter reduction is necessary about every 200 miles in a long digital link to keep the maximum jitter within reasonable limits.

7.6 DETECTION-ERROR PROBABILITY

The signal received at the detector consists of the desired pulse train plus a random channel noise. This can cause error in pulse detection. Consider, for example, the case of polar transmission using a basic pulse $p(t)$ (Fig. 7.24a). This pulse has a peak amplitude A_p. A typical received pulse train is shown in Fig. 7.24b. Pulses are sampled at their peak values. If noise were absent, the sample of the positive pulse (corresponding to **1**) would be A_p and that of the negative pulse (corresponding to **0**) would be $-A_p$.* Because of noise, these samples would be $\pm A_p + n$ where n is the random noise amplitude (see Fig. 7.24b). From the symmetry of the situation, the detection threshold is zero; that is, if the pulse sample value is positive, the digit is detected as **1**; if the sample value is negative, the digit is detected as **0**.

The decision whether **1** or **0** is transmitted could be made readily from the pulse sample, except that n is random, meaning its exact value is unpredictable. It may have a large or a small value and can be negative as well as positive. It is possible that **1** is transmitted, but n at the sampling instant may have a large negative value. This will make the sample value $A_p + n$ small or even negative. On the other hand, if **0** is transmitted, and n has a large positive value at the sampling instant, the sample value $-A_p + n$ can be positive and the digit will be detected wrongly as **1**. This is clear from Fig. 7.24b.

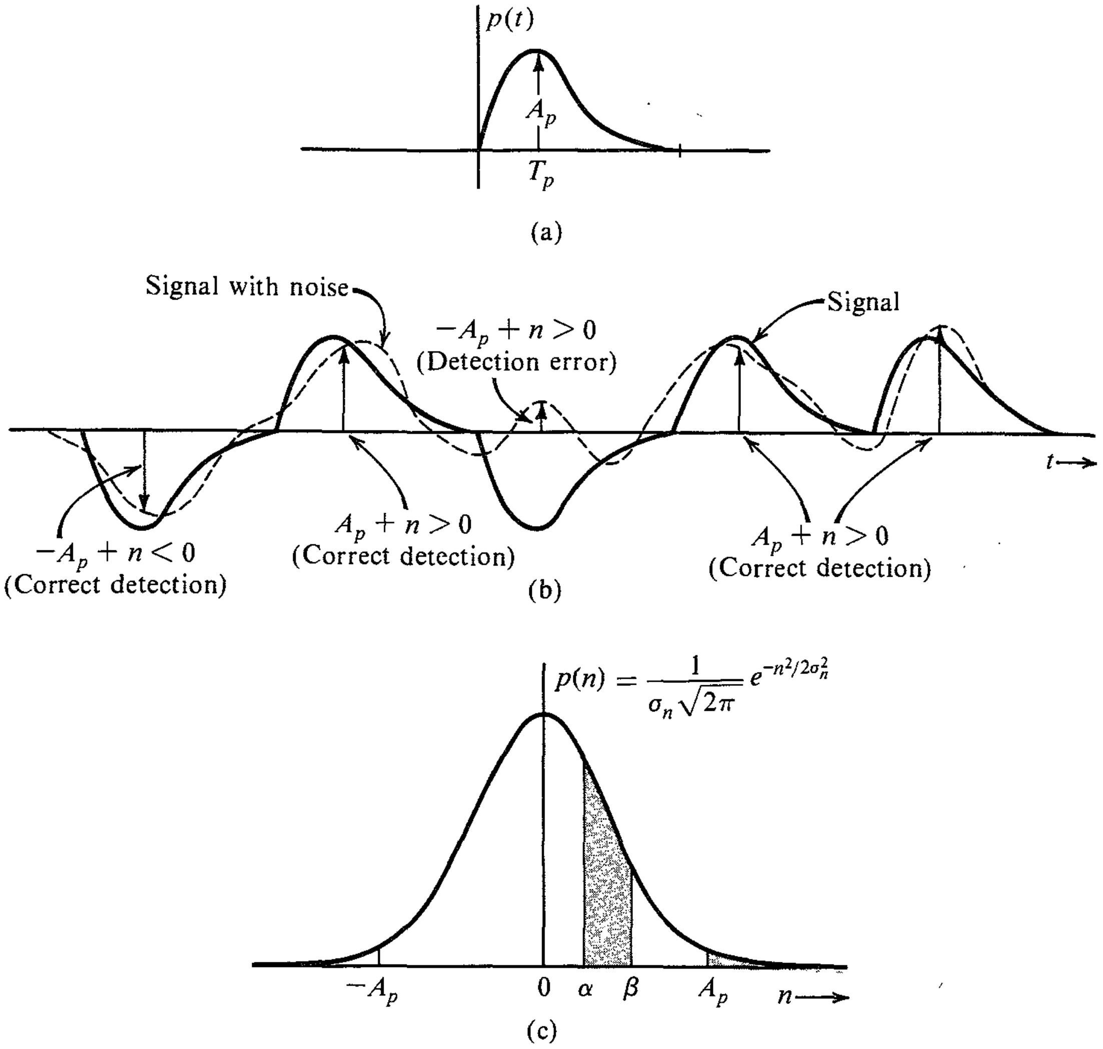

$p(t)$
A_p
T_p
(a)
Signal with noise
$-A_p + n > 0$
(Detection error)
Signal
$t \longrightarrow$
$-A_p + n < 0$
(Correct detection)
$A_p + n > 0$
(Correct detection)
(b)
$A_p + n > 0$
(Correct detection)
$p(n) = \frac{1}{\sigma_n \sqrt{2\pi}} e^{-n^2/2\sigma_n^2}$
$-A_p$
0
α
β
A_p
$n \longrightarrow$
(c)

Figure 7.24 Error probability in threshold detection.

---

* This assumes zero ISI.

Error Probability for Polar Signal

We now compute the likelihood of error (error probability) for a polar signal. The amplitude n of the so-called **gaussian** noise ranges from $-\infty$ to ∞, although the likelihood (or the probability) that n will take on very large values decreases rapidly as $e^{-n^2/2\sigma_n^2}$, where σ_n is the rms value of the noise. Still, occasionally, n can take on large positive or negative values causing detection errors as discussed earlier. When **0** is transmitted, the sample value of the received pulse is $-A_p + n$. If $n > A_p$, the sample value is positive and the digit will be detected wrongly as **1**. If $P(\epsilon|\mathbf{0})$ is the likelihood or the probability of error, given that **0** is transmitted, then

$$P(\epsilon|\mathbf{0}) = \text{likelihood or the probability that } n > A_p \tag{7.46a}$$

Similarly,

$$P(\epsilon|\mathbf{1}) = \text{probability that } n < -A_p \tag{7.46b}$$

The detailed discussion of random signals and error probabilities is deferred to Chapters 10 and 11. The following brief discussion about the basics of probability is adequate for the present need.

The probabilities in Eq. (7.46) can be computed if we know the relative distribution of amplitudes of the random noise. Such distribution is specified by the **probability density function (PDF)** of n. For the case of gaussian noise, the PDF $p(n)$ is given by

$$p(n) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-n^2/2\sigma_n^2} \tag{7.47}$$

where σ_n is the rms value of the noise signal. This PDF, shown in Fig. 7.24c, indicates the relative distribution of values of n. The likelihood (or probability) that an amplitude of n lies in a range (α, β) is given by the area under PDF over the range (α, β), that is,

$$\text{Probability } (\alpha < n < \beta) = \frac{1}{\sigma_n\sqrt{2\pi}} \int_\alpha^\beta e^{-n^2/2\sigma_n^2}\, dn \tag{7.48}$$

Now, $P(\epsilon|\mathbf{0})$ is the probability that $n > A_p$. This is given by the area under $p(n)$ from A_p to ∞ (the shaded tail in Fig. 7.24c). Therefore,

$$P(\epsilon|\mathbf{0}) = \frac{1}{\sigma_n\sqrt{2\pi}} \int_{A_p}^\infty e^{-n^2/2\sigma_n^2}\, dn \tag{7.49a}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{A_p/\sigma_n}^\infty e^{-x^2/2}\, dx \tag{7.49b}$$

This integral cannot be obtained in a closed form. It has been computed numerically and can be found in standard mathematical tables. If we define a function $Q(y)$ as

$$Q(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-x^2/2}\, dx \tag{7.50}$$

Then from Eq. (7.49b) it follows that

$$P(\epsilon|\mathbf{0}) = Q\left(\frac{A_p}{\sigma_n}\right) \tag{7.51a}$$

In a similar way, $P(\epsilon|\mathbf{1})$ is the probability that $n < -A_p$. This is given by the area under $p(n)$ from $-A_p$ to $-\infty$. Because of the symmetry of $p(n)$ about the origin, this area is identical to that from A_p to ∞ (Fig. 7.24c). Therefore, $P(\epsilon|\mathbf{1})$ is the same as $P(\epsilon|\mathbf{0})$, viz., $Q(A_p/\sigma_n)$. If **1** and **0** are equally likely, the average error probability is*

$$P(\epsilon) = \frac{1}{2}[P(\epsilon|\mathbf{0}) + P(\epsilon|\mathbf{1})] = Q\left(\frac{A_p}{\sigma_n}\right) \tag{7.51b}$$

The function $Q(x)$ is plotted in Fig. 10.11d and tabulated in Table 10.2. This function (or its scaled version) also is known in literature as **complementary error function erfc(x)**.

A very good approximation to the Q function is

$$Q(x) \simeq \frac{1}{x\sqrt{2\pi}}\left(1 - \frac{0.7}{x^2}\right)e^{-x^2/2} \qquad x > 2 \tag{7.52}$$

The error in this approximation is just about 1% for $x > 2.15$. The error decreases as x increases.

To get a rough idea of the orders of magnitude, let us consider the peak pulse amplitude A_p to be k times the noise rms value, that is, $A_p = k\sigma_n$. In this case,

$$P(\epsilon|\mathbf{0}) = P(\epsilon|\mathbf{1}) = Q(k)$$

Table 7.1 shows error probabilities for various values of k.

Table 7.1

k	**1**	**2**	**3**	**4**	**5**	**6**
$P(\epsilon\|\mathbf{0})$	0.1587	0.0227	0.00135	3.16×10^{-5}	2.87×10^{-7}	9.9×10^{-9}

The error probability of 10^{-6} means that, on the average, only one out of a million pulses will be detected wrongly. Thus when A_p is five times the noise rms amplitude, the error probability is 2.87×10^{-7}. This means that, on the average, only 1 out of 3,484,320 pulses will be detected wrongly.

Error Probability for On-Off and Bipolar Signals

For the on-off case, we have to distinguish between A_p (presence of a pulse) and 0 (no pulse). The detection threshold is $A_p/2$, that is, if the received sample value is greater than $A_p/2$, we decide in favor of **1**. If the sample value is less than $A_p/2$, we decide in favor of **0**. Thus, following our earlier argument for the polar case, we obtain

$$P(\epsilon|\mathbf{0}) = \text{probability}\left(n > \frac{A_p}{2}\right) = Q\left(\frac{A_p}{2\sigma_n}\right)$$

Similarly,

$$P(\epsilon|\mathbf{1}) = \text{probability}\left(n < -\frac{A_p}{2}\right) = Q\left(\frac{A_p}{2\sigma_n}\right)$$

---

* In this discussion, we have assumed that the occurrence of both symbols **1** and **0** is equally likely. For a general case, if **1** and **0** have probabilities $P(\mathbf{1})$ and $P(\mathbf{0})$, respectively, then

$$P(\epsilon) = P(\mathbf{1})P(\epsilon|\mathbf{1}) + P(\mathbf{0})P(\epsilon|\mathbf{0}) \tag{7.51a}$$

Assuming **1** and **0** equally likely, the average error probability is

$$P(\epsilon) = \frac{1}{2}[P(\epsilon|\mathbf{0}) + P(\epsilon|\mathbf{1})] = Q\left(\frac{A_p}{2\sigma_n}\right) \tag{7.53}$$

In general, if the separation between pulse amplitudes to be distinguished is $2A_p$ (as in the polar case discussed here), the error probability is $Q(A_p/\sigma_n)$. For the on-off case, the separation between amplitudes to be be distinguished is only A_p, and, consequently, the error probability is $Q(A_p/2\sigma_n)$.

In the bipolar case the situation is slightly different because **1** is transmitted by a positive or a negative pulse and **0** is transmitted by no pulse. If the detected sample is in the range $(-A_p/2,\ A_p/2)$, we decide in favor of **0**. If the detected sample is outside this range, we detect it as **1**. Thus,

$$\begin{aligned}
P(\epsilon|\mathbf{0}) &= \text{probability}\left(|n| > \frac{A_p}{2}\right) = \text{Prob}\left(n > \frac{A_p}{2}\right) + \text{Prob}\left(n < -\frac{A_p}{2}\right)\\
&= 2\ \text{probability}\left(n > \frac{A_p}{2}\right)\\
&= 2Q\left(\frac{A_p}{2\sigma_n}\right)
\end{aligned}$$

and

$$\begin{aligned}
P(\epsilon|\mathbf{1}) = {} & \text{probability}\left(n < -\tfrac{A_p}{2}\right) \text{ when positive pulse is used}\\
& \text{or probability}\left(n > \tfrac{A_p}{2}\right) \text{ when negative pulse is used}\\
= {} & Q\left(\frac{A_p}{2\sigma_n}\right)
\end{aligned}$$

Assuming **1** and **0** equally likely, the average error probability is

$$P(\epsilon) = \frac{1}{2}[P(\epsilon|\mathbf{0}) + P(\epsilon|\mathbf{1})] = 1.5\ Q\left(\frac{A_p}{2\sigma_n}\right) \tag{7.54}$$

Thus, $P(\epsilon)$ is 50% higher for the bipolar case than for the on-off case. This may appear as a serious degradation in performance. But because $P(\epsilon)$ decreases exponentially with the signal power, this increase in $P(\epsilon)$ can be compensated by just a little increase in power. For example, to attain a $P(\epsilon)$ of 0.286×10^{-6}, we need $A_p/\sigma_n = 5$ for the polar case because $Q(5) = 0.286 \times 10^{-6}$. To achieve the same $P(\epsilon)$ for the on-off case, we require $A_p/\sigma_n = 10$ [see Eq. (7.53)]. For the bipolar case, to achieve the same $P(\epsilon)$, we require $A_p/\sigma_n = 10.16$ because, from Eq. (7.54), $P(\epsilon) = 1.5Q(5.08) = 0.286 \times 10^{-6}$. This is just a 1.6% increase in the signal amplitude (or a 3.23% increase in the signal power) required over the on-off case. Thus, the performance of the bipolar case is almost the same as that of the on-off case from the point of view of $P(\epsilon)$.

EXAMPLE 7.4

(a) Polar binary pulses are received with peak amplitude $A_p = 1$ mV. The channel noise rms amplitude is 192.3 μV. Threshold detection is used, and **1** and **0** are equally likely. Find the detection error probability.

(b) Find the error probability for **(i)** the on-off case and **(ii)** the bipolar case if pulses of the same shape as in part **(a)** are used, but their amplitudes are adjusted so that the transmitted power is the same as in part **(a)**.

(a) For the polar case,

$$\frac{A_p}{\sigma_n} = \frac{10^{-3}}{192.3(10^{-6})} = 5.2$$

From Table 10.2, we find

$$P(\epsilon) = Q(5.2) = 0.9964 \times 10^{-7}$$

(b) Because half the bits are transmitted by no pulse, there are, on the average, only half as many pulses in the on-off case (compared to the polar). To maintain the same power, we need to double the energy of each pulse in the on-off or the bipolar case (compared to the polar). Now, doubling the pulse energy is accomplished by multiplying the pulse by $\sqrt{2}$. Thus, for on-off A_p is $\sqrt{2}$ times the A_p in the polar case, that is, $A_p = \sqrt{2}(10)^{-3}$. Therefore, from Eq. (7.53),

$$P(\epsilon) = Q\left(\frac{A_p}{2\sigma_n}\right) = Q(3.68) = 1.166 \times 10^{-4}$$

As seen earlier, for a given power, the A_p for both the on-off and the bipolar cases are identical. Hence, from Eq. (7.54),

$$P(\epsilon) = 1.5\, Q\left(\frac{A_p}{2\sigma_n}\right) = 1.749 \times 10^{-4}$$

In practical systems, the principles outlined in this chapter are used to ensure that random channel noise from thermal effects and intersystem cross-talk will cause errors in a negligible percentage of the received pulses. Switching transients, lightning strikes, power line load switching, and other singular events cause very high-level noise pulses of short duration to contaminate the cable pairs that carry digital signals. These pulses, collectively called **impulse noise**, cannot conveniently be engineered away, and they constitute the most prevalent source of errors from the environment outside the digital systems. Errors are virtually never, therefore, found in isolation, but occur in bursts of up to several hundred at a time. To correct error burst, we use special **burst-error-correcting codes**, described in Chapter 16.

7.7 M-ARY COMMUNICATION

Digital communication uses only a finite number of symbols for communication, the minimum number being two (the binary case). Thus far we have restricted ourselves to only the binary case. We shall now briefly discuss some aspects of M-ary communication (communication using M symbols). This subject will be discussed in depth in Chapters 13 and 14.

We can readily show that the information transmitted by each symbol increases with M For example, when $M = 4$ (4-ary or quaternary case), we have four basic symbols, or pulses available for communication (Fig. 7.25a). A sequence of two binary digits can be transmitted by just one 4-ary symbol. This is because a sequence of two binary digits can form only four

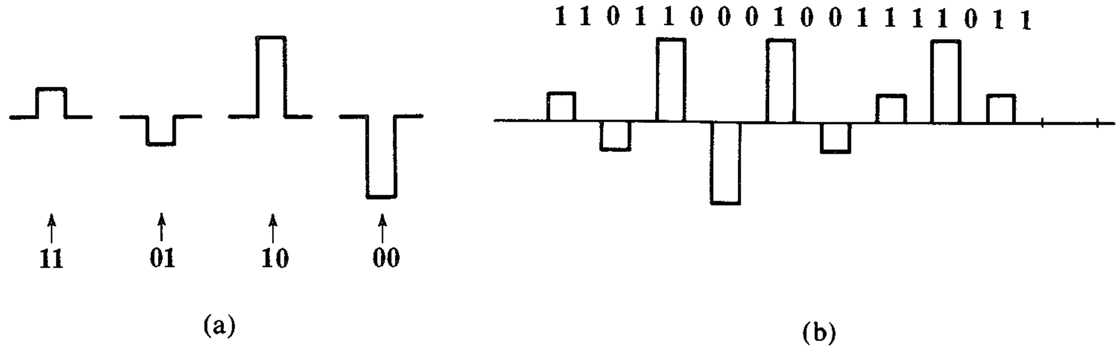


Figure 7.25 4-ary multiamplitude signal.

possible sequences (viz., **11, 10, 01,** and **00**). Because we have four distinct symbols available, we can assign one of the four symbols to each of these combinations (Fig. 7.25a). This signaling (**multiamplitude signaling**) allows us to transmit each pair of binary digits by one 4-ary pulse (Fig. 7.25b). Hence, to transmit n binary digits, we need only $(n/2)$ 4-ary pulses. This means one 4-ary symbol can transmit the information of two binary digits. Also, because three binary digits can form $2 \times 2 \times 2 = 8$ combinations, a group of 3 bits can be transmitted by one 8-ary symbol. Similarly, a group of 4 bits can be transmitted by one 16-ary symbol. In general, the information I_M transmitted by an M-ary symbol is

$$I_M = \log_2 M \text{ binary digits, or bits} \tag{7.55}$$

This means we can increase the rate of information transmission by increasing M. But the transmitted power increases as M, because to have the same noise immunity, the minimum separation between pulse amplitudes should be comparable to that of binary pulses. Therefore, pulse amplitudes increase with M (see Fig. 7.25). It will be shown in Chapter 13 that the transmitted power increases as M^2 (see Prob. 7.7.4). Thus, to increase the rate of communication by a factor of $\log_2 M$, the power required increases as M^2. Because the transmission bandwidth depends only on the pulse rate and not on pulse amplitudes, the bandwidth is independent of M. But if we wish to maintain the same rate of data transmission as in the binary case, we can reduce the transmission bandwidth by a factor of $\log_2 M$ at the cost of increased power.

Although most of the terrestrial digital telephone network uses binary encoding, the subscriber loop portion of the integrated services digital network (ISDN) uses the quaternary code 2B1Q, shown in Fig. 7.25a. It uses NRZ pulses to transmit 160 kbit/s of data using a **baud** rate (pulse rate) of 80 kbit/s. Of the various line codes examined by the ANSI standards committee, 2B1Q provided the greatest baud rate reduction in the noisy and cross-talk-prone local cable plant environment.

Pulse Shaping in the Multiamplitude Case: In this case, we can use the Nyquist criterion pulses because these pulses have zero ISI at the pulse centers, and, therefore, their amplitudes can be correctly detected by sampling at the pulse centers. We can also use the controlled ISI (partial response signaling) for the M-ary case.[9]

Figure 7.25 shows just one possible M-ary scheme (multiamplitude signaling). There are infinite possible ways of structuring M waveforms. For example, we may use M orthogonal pulses $\varphi_1(t)$, $\varphi_2(t)$, ..., $\varphi_M(t)$ with the property

$$\int_0^{T_b} \varphi_i(t)\varphi_j(t)\,dt = \begin{cases} C & i = j \\ 0 & i \neq j \end{cases}$$

Figure 7.26 shows one possible set of M orthogonal signals,

$$\varphi_k(t) = \begin{cases} \sin \frac{2\pi kt}{T_b} & 0 < t < T_b \qquad k = 1,\ 2,\ \ldots,\ M \\ 0 & \text{otherwise} \end{cases}$$

It can be shown that all of these M pulses are mutually orthogonal. Since the highest pulse frequency is M/T_b, the transmission bandwidth in this case is M/T_b. In general it can be shown that the bandwidth of an orthogonal M-ary scheme is M times that of the binary scheme [see Sec. 14.3, Eq. (14.72)]. Therefore, in an M-ary orthogonal scheme, the rate of communication is increased by a factor of $\log_2 M$ at the cost of an increase in transmission bandwidth by a factor of M. For a comparable noise immunity, the transmitted power is practically independent of M in the orthogonal scheme.

In Chapters 13 and 14, we shall discuss several other types of M-ary signaling. The nature of the exchange between the transmission bandwidth and the transmitted power (or SNR) depends on the choice of M-ary scheme. For example, in orthogonal signaling, the transmitted power is practically independent of M, but the transmission bandwidth increases with M. Contrast this to the multiamplitude case, where the transmitted power increases roughly with M^2 and the bandwidth remains constant. Thus, M-ary signaling allows us great flexibility in exchanging signal power (or SNR) with transmission bandwidth. The choice of the appropriate system will depend on the particular circumstances. For instance, it will be appropriate to use multiamplitude signaling if the bandwidth is at a premium (as in telephone lines) and to use orthogonal signaling when power is at a premium (as in space communication). Because of its simplicity, however, binary communication is perhaps the single most important mode of communication in practice today.

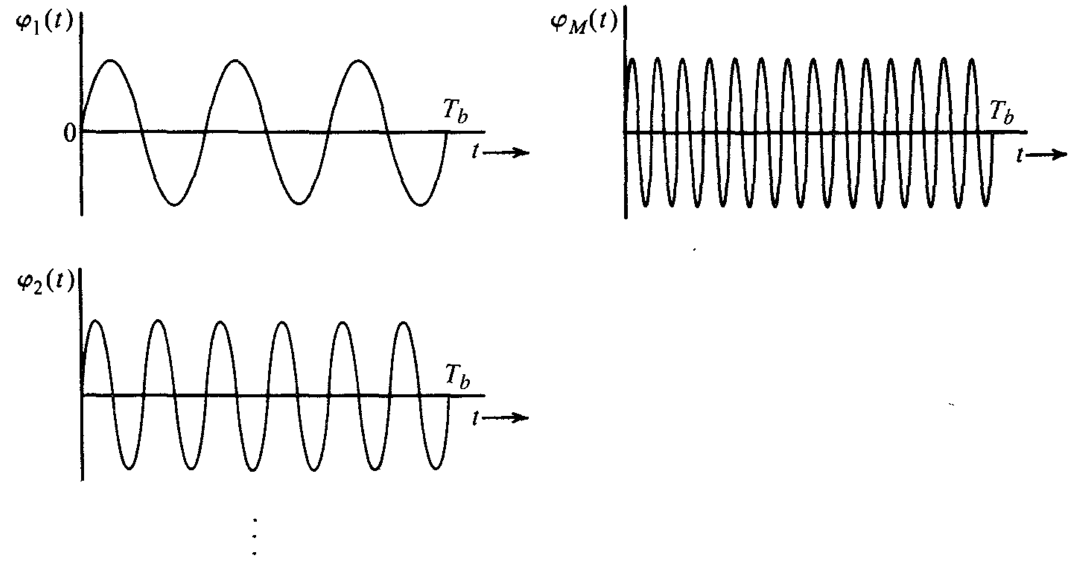


Figure 7.26 M-ary orthogonal pulses.

7.8 DIGITAL CARRIER SYSTEMS

Thus far, we have discussed baseband digital systems, where signals are transmitted directly without any shift in the frequencies of the signal. Because baseband signals have sizable power at low frequencies, they are suitable for transmission over a pair of wires, coaxial cables, or optical fibers. Much of the modern communication is conducted this way. However, baseband signals cannot be transmitted over a radio link or satellites because this would necessitate impracticably large antennas to efficiently radiate the low-frequency spectrum of the signal. Hence, for such a purpose, the signal spectrum must be shifted to a high-frequency range. A spectrum shift to higher frequencies is also required to transmit several messages simultaneously by sharing the large bandwidth of the transmission medium (FDM). As seen in Chapter 3, the spectrum of a signal can be shifted to a higher frequency by modulating a high-frequency sinusoid (carrier) by the baseband signal. Two basic forms of modulation exist: amplitude modulation and angle modulation. In amplitude modulation, the carrier amplitude is varied in proportion to the modulating signal (i.e., the baseband signal). This is shown in Fig. 7.27. An unmodulated carrier $\cos \omega_c t$ is shown in Fig. 7.27a. The on-off baseband signal $m(t)$ (the modulating signal) is shown in Fig. 7.27b. When the carrier amplitude is varied in proportion to $m(t)$, we have the modulated carrier $m(t) \cos \omega_c t$ shown in Fig. 7.27c. Note that the modulated signal is still an on-off signal. This modulation scheme of transmitting binary data is known as **on-off keying (OOK)** or **amplitude-shift keying (ASK)**.

If the baseband signal $m(t)$ were polar (Fig. 7.28a), the corresponding modulated signal $m(t) \cos \omega_c t$ would appear as shown in Fig. 7.28*b*. In this case, if $p(t)$ is the basic pulse, we are transmitting **1** by a pulse $p(t) \cos \omega_c t$ and **0** by $-p(t) \cos \omega_c t = p(t) \cos(\omega_c t + \pi)$. Hence, the two pulses are π radians apart in phase. The information resides in the phase of the pulse. For this reason this scheme is known as **phase-shift keying (PSK)**. Note that the transmission is still polar.

When the data are transmitted by varying the frequency, we have the case of **frequency-shift keying (FSK)**, as shown in Fig. 7.28c. A **0** is transmitted by a pulse of frequency ω_{c_0}, and **1** is transmitted by a pulse of frequency ω_{c_1}. The information about the transmitted data resides in the carrier frequency.

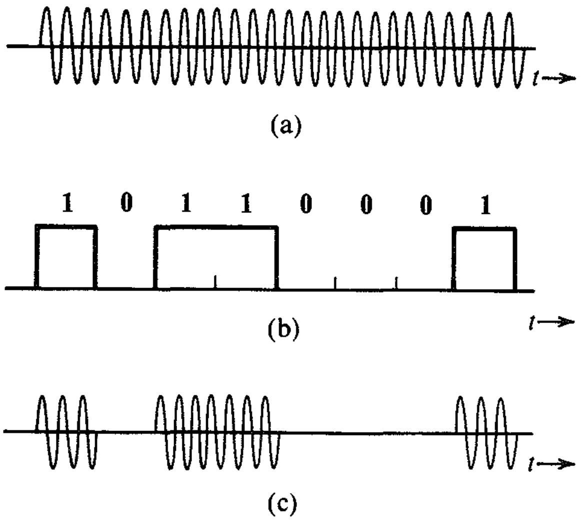


Figure 7.27 (a) Carrier $\cos \omega_c t$. (b) Modulating signal $m(t)$. (c) ASK: modulated signal $m(t) \cos \omega_c t$.

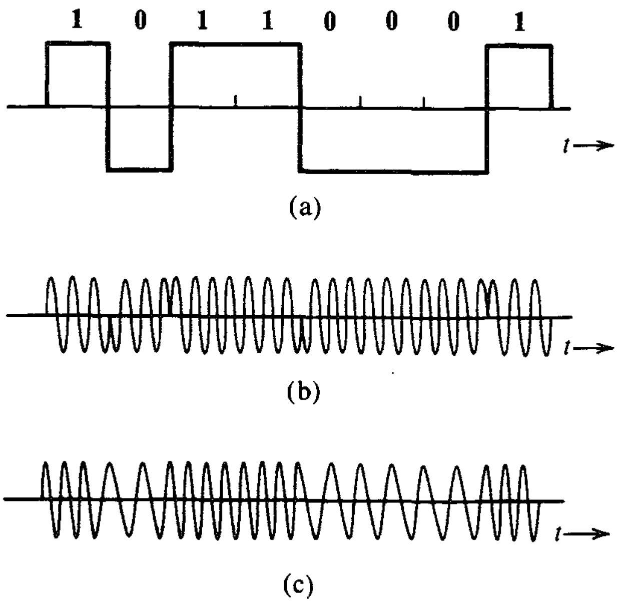


Figure 7.28 (a) Modulating signal $m(t)$. (b) PSK: modulated signal $m(t)\cos\omega_c t$. (c) FSK: modulated signal.

Modulation causes a shift in the baseband signal spectrum [Eq. (3.35)]. The ASK signal in Fig. 7.27c, for example, is $m(t)\cos\omega_c t$, where $m(t)$ is an on-off signal (using a full-width or NRZ pulse). Hence, the PSD of the ASK signal is the same as that of an on-off signal (Fig. 7.7) shifted to $\pm\omega_c$, as shown in Fig. 7.29a.* The PSK signal, on the other hand, is $m(t)\cos\omega_c t$, where $m(t)$ is a polar signal. Therefore, the PSD of a PSK signal is the same as that of the polar baseband signal shifted to $\pm\omega_c$, as shown in Fig. 7.29b. Note that this is the same as the PSD of the ASK minus its discrete components.

The FSK signal may be viewed as a sum of two interleaved ASK signals, one with a modulating frequency ω_{c_0}, and the other with a modulating frequency ω_{c_1}. Hence, the spectrum of FSK is the sum of two ASK spectra at frequencies ω_{c_0} and ω_{c_1}, as shown in Fig. 7.29c. No discrete components appear in this spectrum. It can be shown that by properly choosing ω_{c_0} and ω_{c_1}, discrete components can be eliminated. Note that the bandwidth of FSK is higher than that of ASK or PSK.

We can also modulate bipolar, duobinary, or any other scheme discussed earlier. Use of the basic rectangular pulse in Fig. 7.27 or 7.28 is for the sake of illustration only. In practice, baseband pulses may be specially shaped to eliminate ISI.

As observed earlier, polar signaling is the most power-efficient scheme. The PSK, being polar, requires 3 dB less power than ASK (or FSK) for the same noise immunity, that is, for the same error probability in pulse detection.

Demodulation

Demodulation of digital-modulated signals is similar to that of analog-modulated signals. For example, ASK (Fig. 7.27) can be demodulated coherently (synchronous) or noncoherently (envelope detection). The coherent detector requires mores elaborate equipment and has superior performance, especially when the signal power is low (low SNR). For higher SNR, which is normally the case in practice, the envelope detector performs just as good as the

---

* Note that an on-off signal in Fig. 7.27b uses a full-width rectangular pulse which has no discrete components except at dc. Therefore, the ASK spectrum has a discrete component only at $\omega = \omega_c$.

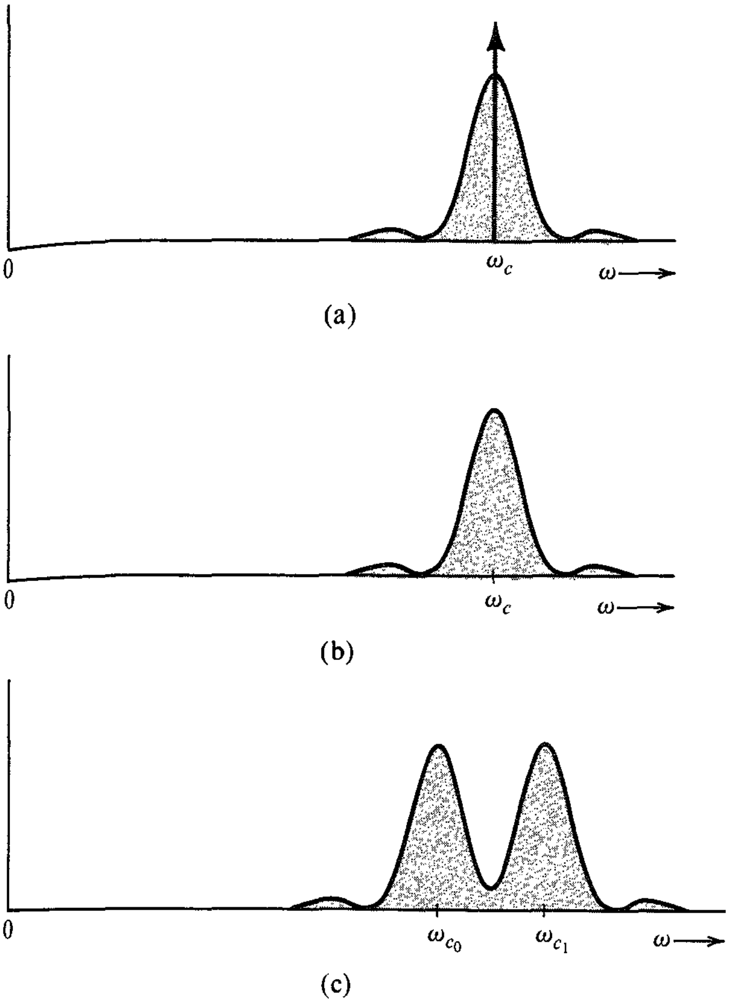


Figure 7.29 PSD of: (a) ASK. (b) PSK. (c) FSK.

synchronous detector. Hence, coherent detection is not used for ASK because it will defeat its very purpose (the simplicity of detection). If we can avail ourselves of a synchronous detector, we might as well use PSK, which is more power efficient than ASK.

In PSK, a **1** is transmitted by a pulse $A \cos \omega_c t$, and a **0** is transmitted by a pulse $-A \cos \omega_c t$ (see Fig. 7.28b). The information in PSK signals therefore resides in the carrier phase. These signals cannot be demodulated noncoherently (envelope detection) because the envelope is the same for both **1** and **0** (see Fig. 7.28b). The coherent detection is similar to that used for analog signals. Methods of carrier acquisition are discussed in Sec. 4.7.

PSK signals may also be demodulated noncoherently using an ingenious method of **differentially coherent PSK (DPSK)**. This technique is facilitated by encoding the data using **differential code** before modulation. In differential code, a **1** is encoded by the same pulse used to encode the previous data bit (no transition) and a **0** is encoded by the negative of the pulse used to encode the previous data bit (transition). This is shown in Fig. 7.30a. Thus, a transition in the received pulse sequence indicates **0** and no transition indicates **1**. The modulated signal consists of pulses $\pm A \cos \omega_c t$. If the data bit is **1**, the present pulse and the previous pulse have the same polarity or phase; both pulses are either $A \cos \omega_c t$ or $-A \cos \omega_c t$. If the data bit is

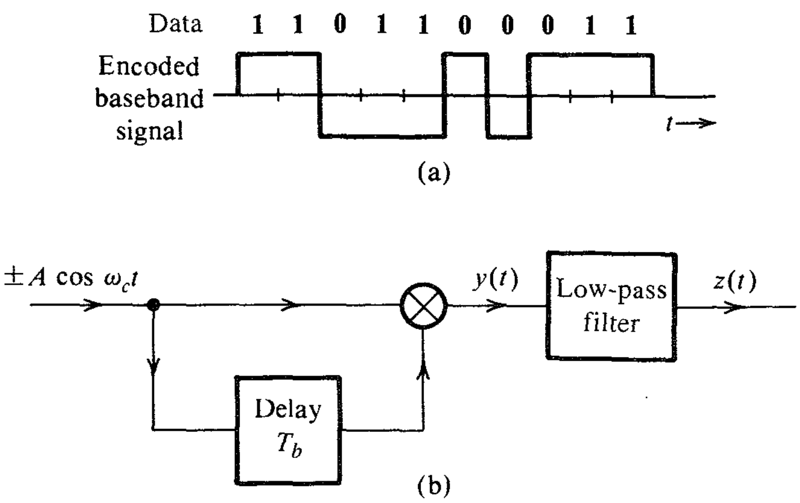

Data 1 1 0 1 1 0 0 0 1 1
Encoded baseband signal
t→
(a)
±A cos ω_c t
y(t)
Low-pass filter
z(t)
Delay
T_b
(b)

Figure 7.30 (a) Differential coding. (b) Differential PSK receiver.

0, the present pulse and the previous pulse are of opposite polarities or phases; if the present pulse is $A\cos\omega_c t$, the previous pulse is $-A\cos\omega_c t$, and vice versa.

In the demodulation of DPSK (Fig. 7.30b), we avoid generation of a local carrier by observing that the received modulated signal itself is a carrier ($\pm A\cos\omega_c t$) with a possible sign ambiguity. For demodulation, in place of the carrier, we use the received signal delayed by T_b (1-bit interval). If the received pulse is identical to the previous pulse, the product $y(t) = A^2\cos^2\omega_c t = (A^2/2)(1+\cos 2\omega_c t)$, and the low-pass filter output $z(t) = A^2/2$. We immediately detect the present bit as **1**. If the received pulse and the previous pulse are of opposite polarity, $y(t) = -A^2\cos^2\omega_c t$ and $z(t) = -A^2/2$, and the present bit is detected as **0**.

The FSK can be viewed as two interleaved ASK signals with carrier frequencies ω_{c_0} and ω_{c_1}, respectively (Fig. 7.28c). Therefore, FSK can be detected coherently or noncoherently.

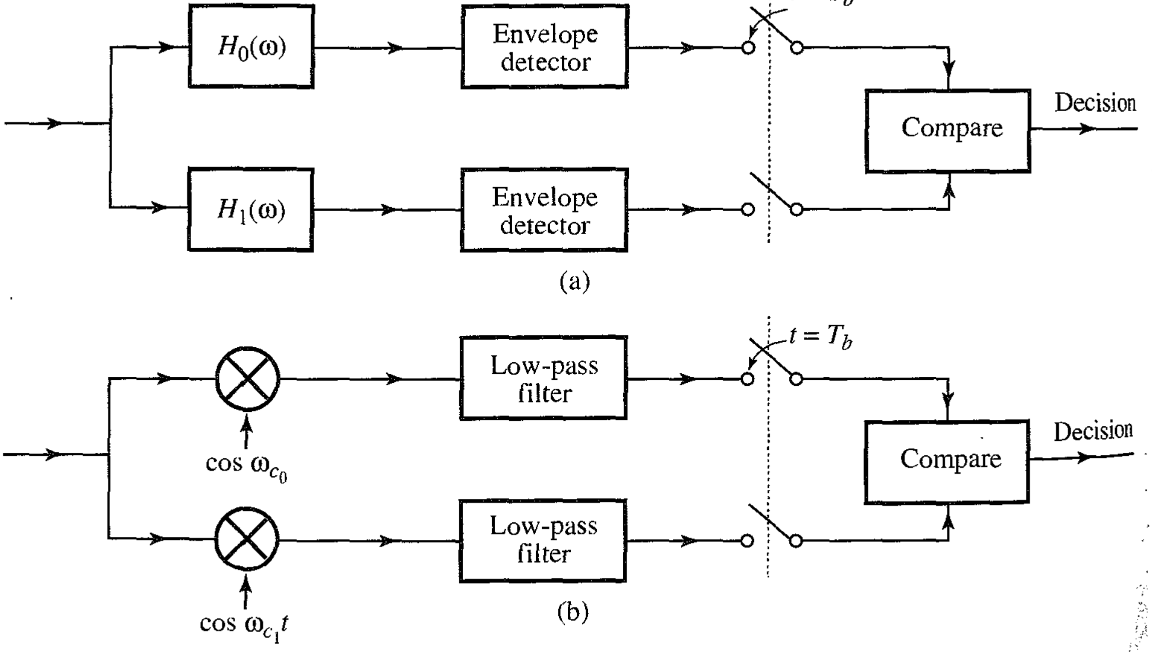

t = T_b
H_0(ω)
Envelope detector
Compare
Decision
H_1(ω)
Envelope detector
(a)
t = T_b
Low-pass filter
cos ω_c0
Compare
Decision
Low-pass filter
cos ω_c1 t
(b)

Figure 7.31 (a) Noncoherent detection of FSK. (b) Coherent detection of FSK.

In noncoherent detection, the incoming signal is applied to a bank of two filters tuned to ω_{c_0} and ω_{c_1}. Each filter is followed by an envelope detector (see Fig. 7.31a). The outputs of the two envelope detectors are sampled and compared. A **0** is transmitted by a pulse of frequency ω_{c_0}, and this pulse will appear at the output of the filter tuned to ω_{c_0}. Practically no signal appears at the output of the filter tuned to ω_{c_1}. Hence, the sample of the envelope detector output following the ω_{c_0} filter will be greater than the sample of the envelope detector output following the ω_{c_1} filter, and the receiver decides that a **0** was transmitted. In the case of a **1**, the situation is reversed.

FSK can also be detected coherently by generating two references of frequencies ω_{c_0} and ω_{c_1}, and demodulating the received signal by two demodulators using the two carriers and then comparing the outputs of the two demodulators, as shown in Fig. 7.31b.

From the point of view of noise immunity, coherent PSK is superior to all other schemes. PSK also requires a smaller bandwidth than FSK (see Fig. 7.29). A quantitative discussion of this topic can be found in Chapter 14.

Digital Signal Transmission Using QAM

Quadrature amplitude modulation (QAM) discussed in Chapter 4 (Fig. 4.14) can be conveniently used for digital signals as well. Figure 7.32a shows the QAM modulator and demodulator. Each of the signals $m_1(t)$ and $m_2(t)$ is a baseband binary polar pulse sequence. These signals are modulated by a carrier of the same frequency but in phase quadrature. Note that both of the modulated signals are PSK signals. For this reason, it is also known as **quadrature PSK (QPSK)**. As seen in Sec. 4.4, we can transmit and receive both of these signals on the same channel, thus doubling the transmission rate.

M-ary QAM

We can increase the transmission rate further by using M-ary QAM.* One practical case with $M = 16$ uses the following 16 pulses (16 symbols):

$$\begin{aligned} p_i(t) &= a_i p(t) \cos \omega_c t + b_i p(t) \sin \omega_c t \\ &= r_i p(t) \cos (\omega_c t - \theta_i) \qquad i = 1, 2, \ldots, 16 \end{aligned}$$

where

$$r_i = \sqrt{a_i^2 + b_i^2} \qquad \text{and} \qquad \theta_i = \tan^{-1} \frac{b_i}{a_i}$$

and $p(t)$ is a properly shaped baseband pulse. The signal $p_i(t)$ can be generated using QAM by letting $m_1(t) = a_i p(t)$ and $m_2(t) = b_i p(t)$. One possible choice of r_i and θ_i for 16 pulses is shown graphically in Fig. 7.32b. The transmitted pulse $p_i(t)$ can take on 16 distinct forms and is, therefore, a 16-ary pulse. Since $M = 16$, each pulse can transmit the information of $\log_2 16 = 4$ binary digits. This can be done as follows: There are 16 possible sequences of four binary digits and there are 16 combinations (a_i, b_i) in Fig. 7.32b. Thus, every possible 4-bit sequence is transmitted by a particular (a_i, b_i) or (r_i, θ_i). Therefore, one signal pulse $r_i p(t) \cos (\omega_c t - \theta_i)$ transmits 4 bits. The bit rate is quadrupled without increasing the bandwidth. The transmission rate can be increased further by increasing the value of M.

Modulation as well as demodulation can be performed by using the system in Fig. 7.32a. The inputs are $m_1(t) = a_i p(t)$ and $m_2(t) = b_i p(t)$. The two outputs at the demodulator are

---

* The M-ary QAM discussed here is also called **amplitude phase keying (APK)**.

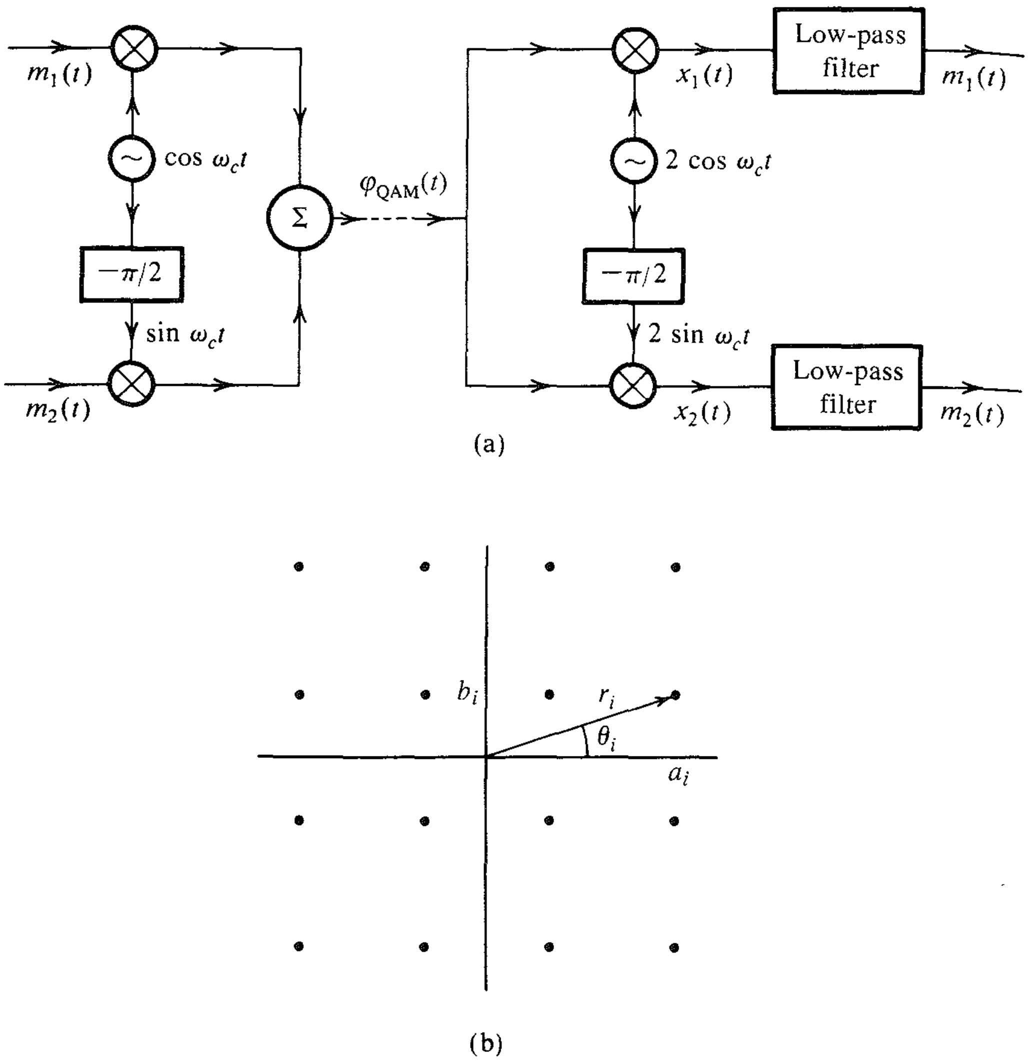

$m_1(t)$
$\cos \omega_c t$
$-\pi/2$
$\sin \omega_c t$
$m_2(t)$
Σ
$\varphi_{\text{QAM}}(t)$
$x_1(t)$
$2 \cos \omega_c t$
$-\pi/2$
$2 \sin \omega_c t$
$x_2(t)$
Low-pass filter
Low-pass filter
$m_1(t)$
$m_2(t)$
(a)
b_i
r_i
θ_i
a_i
(b)

Figure 7.32 (a) QAM or quadrature multiplexing. (b) 16-point QAM ($M = 16$).

$a_i p(t)$ and $b_i p(t)$. From the knowledge of $(a_i,\ b_i)$, we can determine the four transmitted bits. Further analysis of 16-ary QAM on a noisy channel is carried out in Example 14.3.

Such a QAM scheme is used on telephone lines for data transmission. At each end of the telephone line, we need a modulator and a demodulator to transmit as well as to receive data. The two devices, **mo**dulator and **dem**odulator, are usually packaged in one unit called a **modem**.

7.9 DIGITAL MULTIPLEXING

Several low-bit-rate signals can be multiplexed, or combined, to form one high-bit-rate signal to be transmitted over a high-frequency medium. Because the medium is time-shared by various incoming signals, this is a case of time-division multiplexing (TDM). The signals from various incoming channels, or tributaries, may be of such diverse nature as a digitized voice signal

(PCM), a computer output, telemetry data, a digital facsimile, and so on. The bit rates of the various tributaries need not be the same.

To begin with, consider the case of all tributaries with identical bit rates. Multiplexing can be done on a bit-by-bit basis (known as **bit** or **digit interleaving**), as shown in Fig. 7.33a, or on a word-by-word basis (known as **byte** or **word interleaving**). Figure 7.33b shows the interleaving of words, or bytes, formed by 4 bits. The North American digital hierarchy uses bit interleaving (except at the lowest level), where bits are taken one at a time from the various signals to be multiplexed. Byte interleaving, used in building the DS1 signal and SONET-formatted signals (described in the next chapter), involves inserting bytes (8 bits) in succession from the channels to be multiplexed.

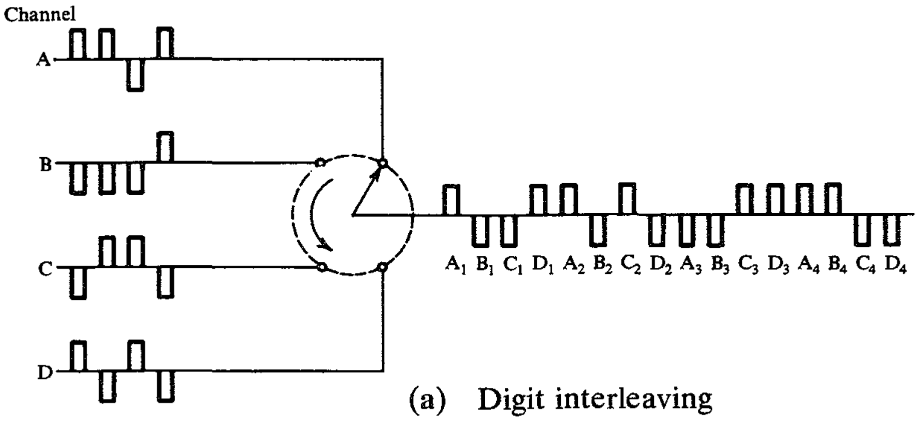
The T1 carrier, discussed in Sec. 6.2.4, uses 8-bit word interleaving. When the bit rates of incoming channels are not identical, the high-bit-rate channel is allocated proportionately more slots. Figure 7.33c and d shows four-channel multiplexing consisting of three channels (B, C, and D) of identical bit rate R and one channel (channel A) with a bit rate of $3R$. Similar results can be attained by combining words of different lengths. It is evident that the minimum length of the multiplex frame must be a multiple of the lowest common multiple of the incoming channel bit rates, and, hence, this type of scheme is practical only when some fairly simple relationship exists among these rates. The case of completely asynchronous channels is discussed later.

At the receiving terminal, the incoming digit stream must be divided and distributed to the appropriate output channel. For this purpose, the receiving terminal must be able to identify each bit correctly. This requires the receiving system to uniquely synchronize in time with the beginning of each frame, with each slot in a frame, and with each bit within a slot. This is accomplished by adding framing and synchronization bits to the data bits. These bits are part of the so-called **overhead bits.**

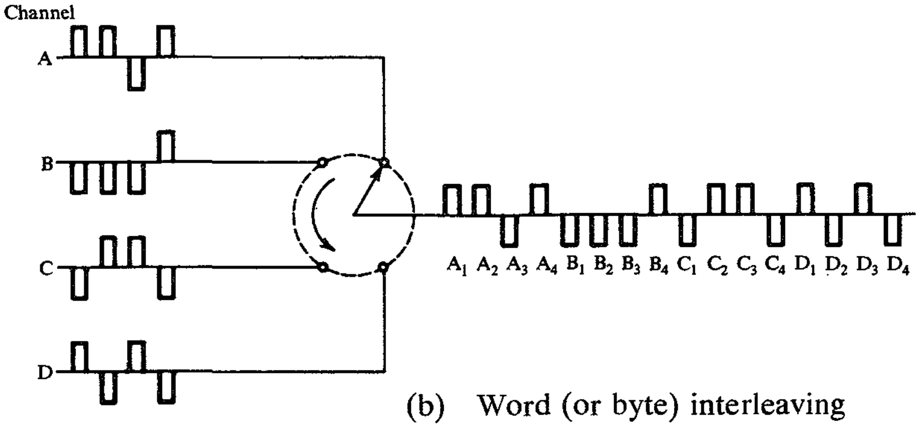
Signal Format

Figure 7.34 illustrates a typical format, that of the DM1/2 multiplexer. We have here bit-by-bit interleaving of four channels, each at a rate of 1.544 Mbit/s. The main frame (multiframe) consists of four subframes. Each subframe has six overhead bits. For example, subframe 1 (first line in Fig. 7.34) has overhead bits M_0, C_A, F_0, C_A, C_A, and F_1. In between these overhead bits are 48 interleaved data bits from the four channels (twelve data bits from each channel). We begin with overhead bit M_0, followed by 48 multiplexed data bits, then add a second overhead bit C_A followed by the next 48 multiplexed bits, and so on. Thus, there are a total of $48 \times 6 \times 4 = 1152$ data bits and $6 \times 4 = 24$ overhead bits making a total of 1176 bits per frame. The efficiency is $1152/1176 \simeq 98\%$. The overhead bits with subscript 0 are always **0** and those with subscript 1 are always **1**. Thus, M_0, F_0 are all **0**'s and M_1 and F_1 are all **1**'s. The F digits are periodic **010101** . . . and provide the main framing pattern. The multiplexer uses this to synchronize on the frame. After locking onto this pattern, the demultiplexer searches for the **0111** pattern formed by overhead bits $M_0M_1M_1M_1$. This further identifies the four subframes, each corresponding to a line in Fig. 7.34. It is possible, although unlikely, that signal bits may also have a pattern **101010** . . . . The receiver could lock onto this wrong sequence. The presence of $M_0M_1M_1M_1$ provides verification of the genuine $F_0F_1F_0F_1$ sequence. The C bits are used to transmit additional information about bit stuffing, as will be discussed later.

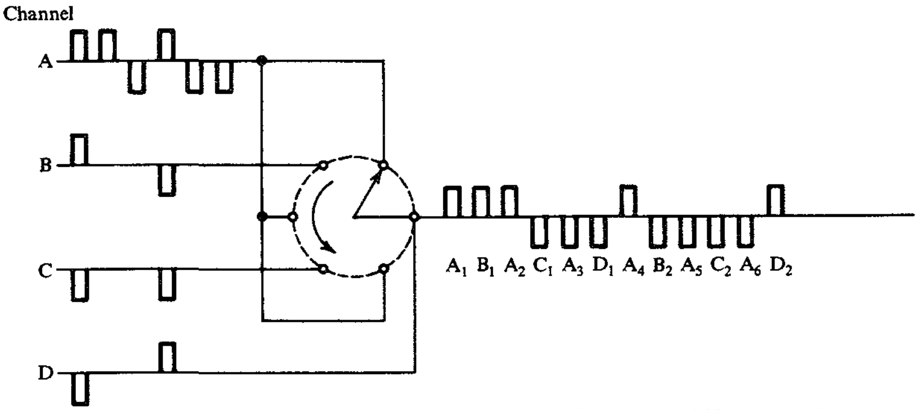
An example of a word interleaving multiplexer (used for PCM) was discussed in Sec. 6.2.4.


Channel
A
B
C
D
A_1 B_1 C_1 D_1 A_2 B_2 C_2 D_2 A_3 B_3 C_3 D_3 A_4 B_4 C_4 D_4

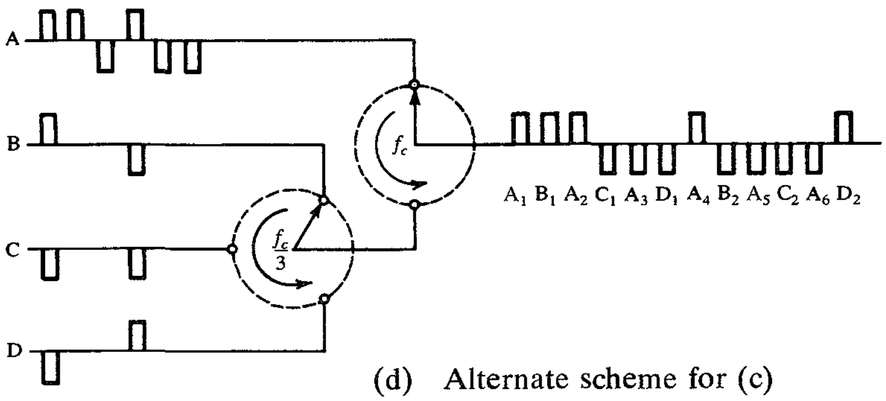
(a) Digit interleaving


Channel
A
B
C
D
A_1 A_2 A_3 A_4 B_1 B_2 B_3 B_4 C_1 C_2 C_3 C_4 D_1 D_2 D_3 D_4

(b) Word (or byte) interleaving


Channel
A
B
C
D
A_1 B_1 A_2 C_1 A_3 D_1 A_4 B_2 A_5 C_2 A_6 D_2

(c) Interleaving channel having different bit rate


A
B
C
D
f_c
$\frac{f_c}{3}$
A_1 B_1 A_2 C_1 A_3 D_1 A_4 B_2 A_5 C_2 A_6 D_2

(d) Alternate scheme for (c)

Figure 7.33 Time-division multiplexing of digital signals.

In the majority of cases not all incoming channels are active all the time. Some of them transmit data and some are idle. This means the system is underutilized. We can, therefore, accept more input channels to take advantage of the fact that some channel will be inactive at any given time. This obviously involves much more complicated switching operations, and also requires rather careful system planning. In any random traffic situation we cannot guarantee that the number of transmission channels demanded will not exceed the number available, but by taking account of the statistics of the signal sources, it is possible to ensure that the probability of this occurring becomes acceptably low. Multiplex structures of this type have been developed for satellite systems and are known as **time-division multiple-access (TDMA) systems**.

In TDMA systems used for telephony, the design parameters are chosen so that any overload condition only lasts for a fraction of a second, which leads to acceptable performance for speech communication. For other types of data and telegraphy, transmission delays are unimportant. Hence, in overload condition, the incoming data can be stored and transmitted later.

Asynchronous Channels and Bit Stuffing

In the preceding discussion, we assumed synchronization between all the incoming channels and the multiplexer. This is difficult even when all the channels are nominally at the same rate. For example, consider a 1000-km coaxial cable carrying 2×10^8 pulses per second. Assuming the nominal propagation speed in the cable to be 2×10^8 m/s, it takes 1/200 second of transit time and 1 million pulses will be in transit. If the cable temperature increases by 1°F, the propagation velocity will increase by about 0.01%. This will cause the pulses in transit to arrive sooner, thus causing a temporary increase in the rate of pulses received. Because the extra pulses cannot be accommodated in the multiplexer, they must be temporarily stored at the receiver. If the cable temperature drops, the rate of received pulses will drop, and the multiplexer will have vacant slots with no data. These slots need to be stuffed with dummy digits (**pulse stuffing**).

DS1 signals in North American networks are often generated by crystal oscillators in individual channel banks or other digital terminal equipment. Although the oscillators are quite stable, they will not oscillate at exactly the same frequency, leading to another cause of asynchronicity in the network.

This shows that even in synchronously multiplexed systems, the data are rarely received at a synchronous rate. We always need a storage (known as an **elastic store**) and pulse stuffing (also known as **justification**) to accommodate such a situation. Obviously, this method of an elastic store and pulse stuffing will work even when the channels are asynchronous.

Three variants of the pulse stuffing scheme exist: (1) positive pulse stuffing, (2) negative pulse stuffing, and (3) positive/negative pulse stuffing. In positive pulse stuffing, the multiplexer rate is higher than that required to accommodate all incoming tributaries at their maximum

M_0	[48]	C_A	[48]	F_0	[48]	C_A	[48]	C_A	[48]	F_1	[48]
M_1	[48]	C_B	[48]	F_0	[48]	C_B	[48]	C_B	[48]	F_1	[48]
M_1	[48]	C_C	[48]	F_0	[48]	C_C	[48]	C_C	[48]	F_1	[48]
M_1	[48]	C_D	[48]	F_0	[48]	C_D	[48]	C_D	[48]	F_1	[48]

Figure 7.34 DM1/2 multiplexer format.

rate. Hence, the time slots in the multiplexed signal will become available at a rate exceeding that of the incoming data so that the tributary data will tend to lag (Fig. 7.35). At some stage, the system will decide that this lag has become great enough to require pulse stuffing. The information about the stuffed-pulse positions is transmitted through overhead bits. From the overhead bits, the receiver knows the stuffed-pulse position and eliminates that pulse.

Negative pulse stuffing is a complement of positive pulse stuffing. The time slots in the multiplexed signal now appear at a slightly slower rate than those of the tributaries so that some of the tributary pulses cannot be accommodated in the multiplexed signal. The information about the left-out pulse and its position is transmitted through overhead bits. The positive/negative pulse stuffing is a combination of these two schemes. The nominal rate of the multiplexer is equal to the nominal rate required to accommodate all incoming channels. Hence, we may need positive pulse stuffing at some times and negative stuffing at others. All this information is sent through overhead bits.

The C digits in Fig. 7.34 are used to transmit stuffing information. Only one stuffed bit per input channel is allowed per frame. This is sufficient to accommodate expected variations in the input signal rate. The bits C_A convey information about stuffing in channel A and bits C_B convey information about stuffing in channel B, and so on. The insertion of any stuffed pulse in any one subframe is denoted by setting all the three C's in that line to **1**. No stuffing is indicated by using **0**'s for all the three C's. If a bit has been stuffed, the location of the stuffed bit is the first information bit associated with the immediate channel following the F_1 bit, that is, the first such bit in the last 48-bit sequence in that subframe. For the first subframe, the stuffed bit will immediately follow the F_1 bit. For the second subframe, the stuffed bit will be the second bit following the F_1 bit, and so on.

Digital Hierarchy

The following is the digital hierarchy developed by the Bell System and currently encoded in ANSI standards for telecommunications (Fig. 7.36). This North American hierarchy is used in North America and Japan.

Two major classes of multiplexers are used in practice. The first category is used for combining low-data-rate channels. It multiplexes channels of rates of up to 9600 bit/s into a signal of data rates of up to 64 kbit/s. The multiplexed signal, called **digital signal level 0 (DS0)** in North American hierarchy, is eventually transmitted over a voice-grade channel. The second class of multiplexers is at a much higher bit rate. There are four orders, or levels, of multiplexing. The first level is the **T1 multiplexer** or **channel bank**, consisting of 24 channels of 64 kbit each. The output of this multiplexer is a **DS1 (digital level 1)** signal at a rate of 1.544 Mbit/s. Four DS1 signals are multiplexed by a DM1/2 multiplexer to yield a DS2 signal at a rate of 6.312 Mbit/s. Seven DS2 signals are multiplexed by a DM2/3 multiplexer to yield a DS3

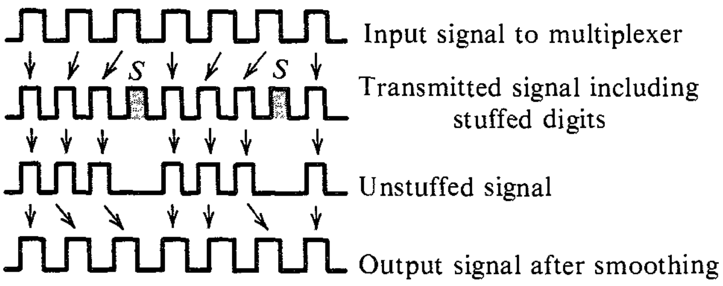

Input signal to multiplexer
S
S
Transmitted signal including stuffed digits
Unstuffed signal
Output signal after smoothing

Figure 7.35 Pulse stuffing.

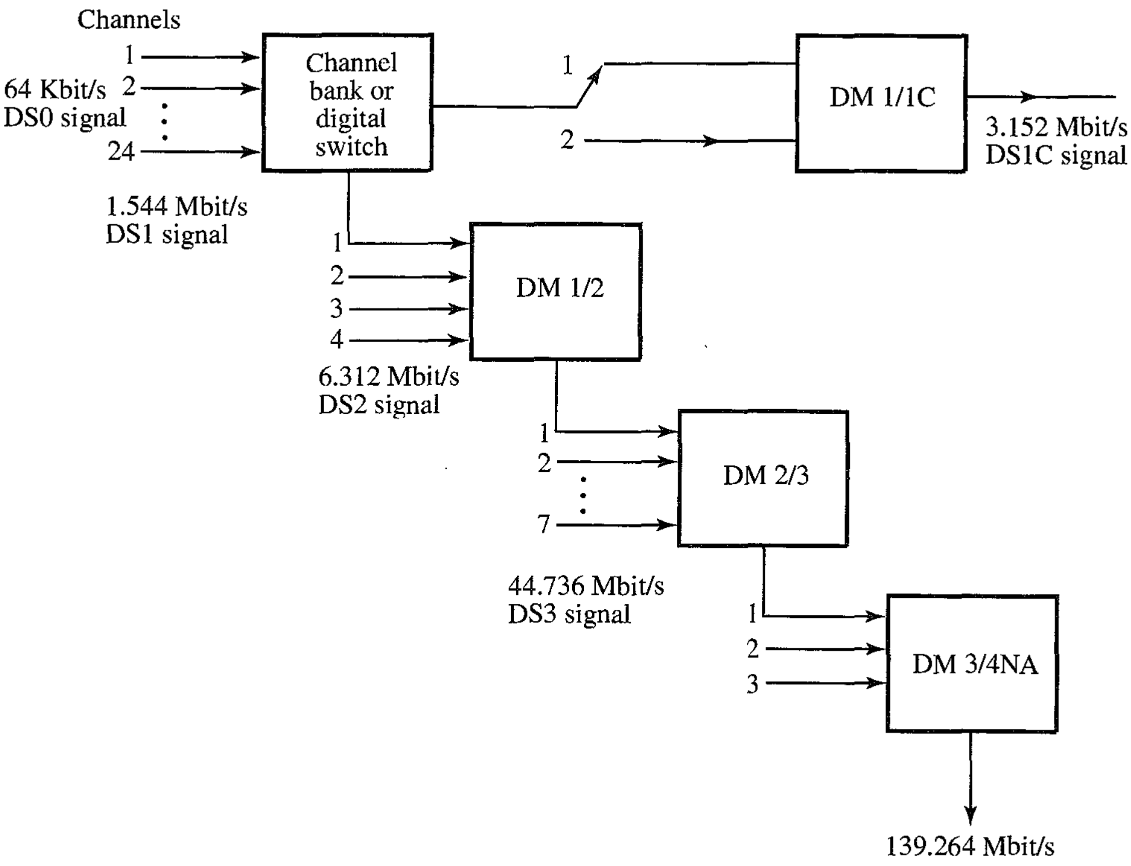

Channels
1
64 Kbit/s 2
DS0 signal
24
Channel bank or digital switch
1.544 Mbit/s
DS1 signal
1
2
DM 1/1C
3.152 Mbit/s
DS1C signal
1
2
3
4
DM 1/2
6.312 Mbit/s
DS2 signal
1
2
7
DM 2/3
44.736 Mbit/s
DS3 signal
1
2
3
DM 3/4NA
139.264 Mbit/s

Figure 7.36 North American digital hierarchy (AT&T system).

signal at a rate of 44.736 Mbit/s. Finally, three DS3 signals are multiplexed by a DM3/4NA multiplexer to yield a DS4NA signal at a rate of 139.264 Mbit/s. There is also a lower rate multiplexing hierarchy, known as the **digital data system (DDS)**, which provides standards for multiplexing digital signals with rates as low as 2.4 kbit/s into a DS0 signal for transmission through the network.

The multiplexer DM 1/1C is useful for channeling paired cable plant but it cannot be multiplexed into higher level signals. Although DS1C is a "dead end" rate, which cannot be multiplexed to higher levels, it has proved useful for channeling interoffice cable pairs with more channels than can be carried by a comparable DS1 system. Although lightwave interoffice transmission has diminished the importance of DS1C, there are still a number of T1C lines in service in the Northern American network.

The inputs to the T1 multiplexer need not be restricted only to digitized voice channels. Any digital signal of 64 kbit/s of appropriate format can be transmitted. The case of the higher levels is similar. For example, all the incoming channels of the DM1/2 multiplexer need not be DS1 signals obtained by multiplexing 24 channels of 64 kbit/s each. Some of them may be 1.544 Mbit/s digital signals of appropriate format, and so on.

In Europe and the rest of the world, another hierarchy, recommended by the CCITT (Consultative Committee on International Telephony and Telegraphy) as an international standard, has been adopted. This hierarchy, based on the lowest level PCM international standard of 2.048 Mbit/s (30 channels), is shown in Fig. 7.37.

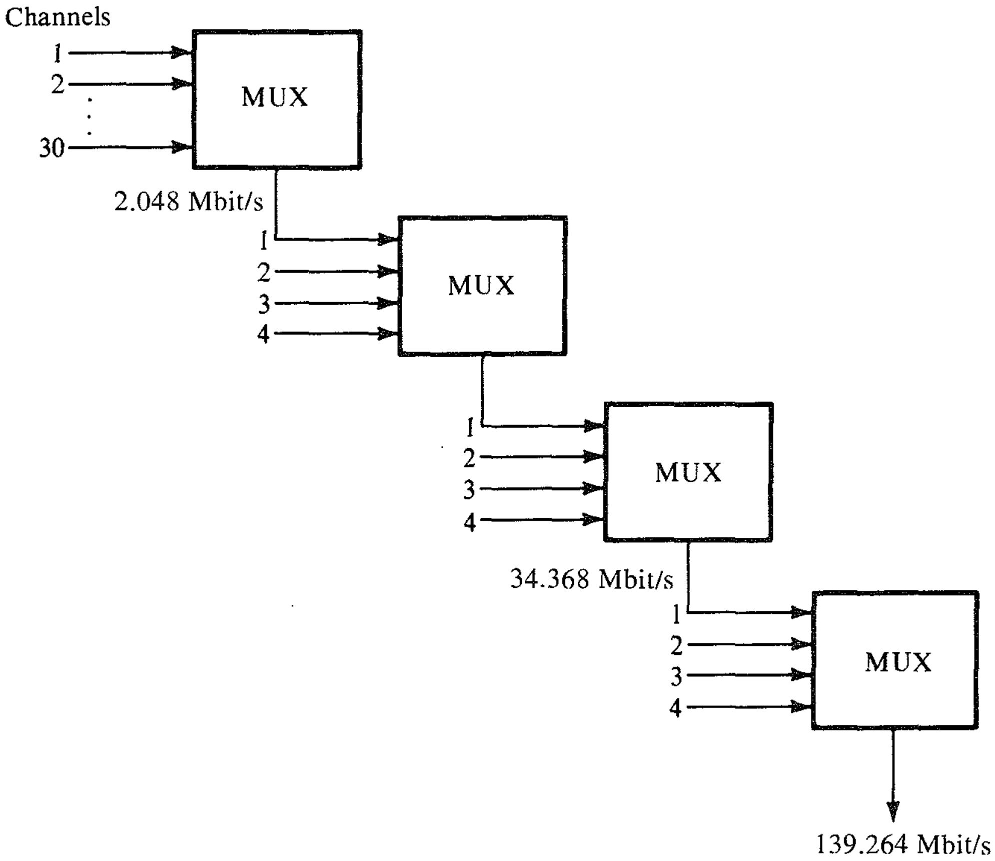


Figure 7.37 Digital hierarchy, CCITT recommendation.

REFERENCES

1. A. Lender, "Duobinary Technique for High Speed Data Transmission," *IEEE Trans. Commun. Electron.*, vol. CE-82, pp. 214–218, May 1963.
2. A. Lender, "Correlative Level Coding for Binary-Data Transmission," *IEEE Spectrum*, vol. 3, pp. 104–115, Feb. 1966.
3. D. W. Davis and D. L. A. Barber, *Communication Networks for Computers*, Wiley, New York, 1973.
4. P. Bylanski and D. G. W. Ingram, *Digital Transmission Systems*, Peter Peregrinus Ltd., Herts., England, 1976.
5. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *AIEE Trans.*, vol. 47, p. 817, Apr. 1928.
6. E. D. Sunde, *Communication Systems Engineering Technology*, Wiley, New York, 1969.
7. R. W. Lucky and H. R. Rudin, "Generalized Automatic Equalization for Communication Channels," *IEEE Int. Comm. Conf.*, vol. 22, 1966.
8. R. W. Lucky, J. Salz, and E. J. Weldon, Jr., *Principles of Data Communication*, McGraw-Hill, New York, 1968.
9. A. Lender, 7 in *Digital Communications: Microwave Applications*, K. Feher, Ed., Prentice Hall, Englewood Cliffs, NJ, 1981, chap. 7.

7.2-1 (a) Find PSDs for polar, on-off, and bipolar signaling, where $p(t)$ is a full-width rectangular pulse, that is, $p(t) = \text{rect}\,(t/T_b)$.

(b) Sketch roughly these PSDs and find their bandwidths. For each case, compare the bandwidth to the case where $p(t)$ is a half-width rectangular pulse.

7.2-2 (a) A random binary data sequence **100110**$\cdots$ is transmitted using a Manchester (split-phase) line code with the pulse $p(t)$ shown in Fig. 7.6a. Sketch the waveform $y(t)$.

(b) Derive $S_y(\omega)$, the PSD of a Manchester (split-phase) signal in part (a) assuming **1** and **0** equally likely. Roughly sketch this PSD and find its bandwidth.

7.2-3 Derive the PSD for a binary signal using differential code with half-width rectangular pulses (see Fig. 7.17). Determine the PSD $S_y(\omega)$.

7.2-4 The duobinary line coding proposed by Lender is also ternary like bipolar, but requires only half the bandwidth of bipolar. In practice, duobinary coding is indirectly realized using a special pulse shape, as discussed in Sec. 7.3. In this code, a **0** is transmitted by no pulse, and a **1** is transmitted by a pulse $p(t)$ or $-p(t)$ using the following rule. A **1** is encoded by the same pulse as that used for the previous **1**, if there is an even number of **0**'s between them. It is encoded by a pulse of opposite polarity if there is an odd number of **0**'s between them. The number 0 is considered an even number. Like bipolar, this code also has a single error detection capability, because correct reception implies that between successive pulses of the same polarity, an even number of **0**'s must occur, and between successive pulses of opposite polarity, an odd number of **0**'s must occur.

(a) Using a half-width pulse $p(t)$, sketch the duobinary signal $y(t)$ for the random binary sequence **1110001101001010**$\cdots$.

(b) Determine R_0, R_1, and R_2 for this code. Assume (or you may show if you like) that $R_n = 0$ for all $n > 2$. Find and sketch the PSD for this line code (assuming half-width pulse). Show that its bandwidth is $R_b/2$ Hz, half that of bipolar.

7.3-1 Data at a rate of 6 kbit/s is to be transmitted over a leased line of bandwidth 4 kHz using Nyquist criterion pulses. Determine the maximum value of the roll-off factor r that can be used.

7.3-2 In a certain telemetry system, there are eight analog measurements, each of bandwidth 2 kHz. Samples of these signals are time-division multiplexed, quantized, and binary coded. The error in sample amplitudes cannot be greater than 1% of the peak amplitude.

(a) Determine L, the number of quantization levels.

(b) Find the transmission bandwidth B_T if Nyquist criterion pulses with roll-off factor $r = 0.2$ are used. The sampling rate must be at least 25% above the Nyquist rate.

7.3-3 A leased telephone of bandwidth 3 kHz is used to transmit binary data. Calculate the data rate (in bits per second) that can be transmitted if we use:

(a) Polar signal with rectangular half-width pulses.

(b) Polar signal with rectangular full-width pulses.

(c) Polar signal using Nyquist criterion pulses of $r = 0.25$.

(d) Bipolar signal with rectangular half-width pulses.

(e) Bipolar signal with rectangular full-width pulses.

7.3-4 The Fourier transform $P(\omega)$ of the basic pulse $p(t)$ used in a certain binary communication system is shown in Fig. P7.3-4.

(a) From the shape of $P(\omega)$, explain if this pulse satisfies the Nyquist criterion.

(b) Find $p(t)$ and verify that this pulse does (or does not) satisfy the Nyquist criterion.

(c) If the pulse does satisfy the Nyquist criterion, what is the transmission rate (in bits per second) and what is the roll-off factor?

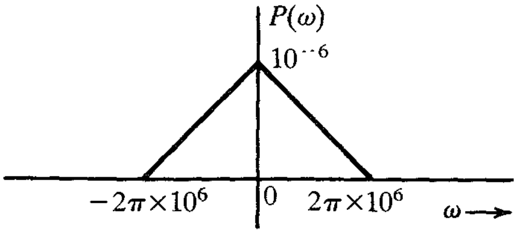


Figure P7.3-4

7.3-5 A pulse $p(t)$ whose spectrum $P(\omega)$ is shown in Fig. P7.3-5 satisfies the Nyquist criterion. If $f_1 = 0.8$ MHz and $f_2 = 1.2$ MHz, determine the maximum rate at which binary data can be transmitted by this pulse using the Nyquist criterion. What is the roll-off factor?

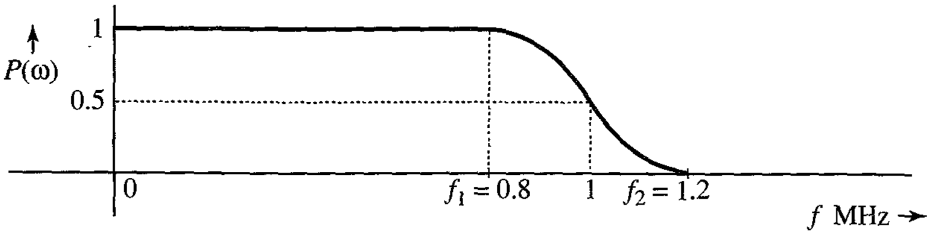


Figure P7.3-5

7.3-6 Binary data at a rate of 1 Mbit/s is to be transmitted using Nyquist criterion pulses with $P(\omega)$ shown in Fig. P7.3-5. The frequencies f_1 and f_2 (in hertz) of this spectrum are adjustable. The channel available for the transmission of this data has a bandwidth of 700 kHz. Determine f_1 and f_2 and the roll-off factor.

7.3-7 Show that the inverse Fourier transform of $P(\omega)$ in Eqs. (7.34) is the pulse $p(t)$ given in Eq. (7.35). *Hint*: Use Eq. (3.32) to find the inverse transform of $P(\omega)$ in Eq. (7.34a) and express sinc (x) in the form sin x/x.

7.3-8 Show that the inverse Fourier transform of $P(\omega)$ in Eq. (7.38) is indeed the duobinary pulse $p(t)$ given in Eq. (7.37). *Hint*: Use Eq. (3.32) to find the inverse transform of $P(\omega)$ in Eq. (7.38) and express sinc(x) in the form sin x/x.

7.3-9 Show that there exists one (and only one) pulse $p(t)$ of bandwidth $R_b/2$ Hz that satisfies the criterion of duobinary pulse [Eq. (7.36)]. Show that this pulse is given by

$$p(t) = \{\text{sinc}\,(\pi R_b t) + \text{sinc}\,[\pi R_b(t - T_b)]\} = \frac{\sin(\pi R_b t)}{\pi R_b t(1 - R_b t)}$$

and its Fourier transform is $P(\omega)$ given in Eq. (7.38). *Hint*: For a pulse of bandwidth $R_b/2$, the Nyquist interval is $1/R_b = T_b$, and conditions (7.36) give the Nyquist sample values at $t = \pm nT_b$. Use the interpolation formula [Eq. (6.10)] with $B = R_b/2$, $T_s = T_b$ to construct $p(t)$. In determining $P(\omega)$, recognize that $(1 + e^{-j\omega T_b}) = e^{-j\omega\frac{T_b}{2}}(e^{j\omega\frac{T_b}{2}} + e^{-j\omega\frac{T_b}{2}})$.

7.3-10. In a binary data transmission using duobinary pulses, sample values were read as follows:

$$1\,2\,0\,-2\,-2\,0\,0\,-2\,0\,2\,0\,0\,2\,0\,0\,0\,-2$$

(a) Explain if there is any error in detection.

(b) If there is no detection error, determine the received bit sequence.

7.3-11 In a binary data transmission using duobinary pulses, sample values of the received pulses were read as follows:

$$1\,2\,0\,0\,0\,-2\,0\,0\,-2\,0\,2\,0\,0\,-2\,0\,2\,2\,0\,-2$$

(a) Explain if there is any error in detection.

(b) Can you guess the correct transmitted digit sequence? There is more than one possible correct sequence. Give as many as possible correct sequences, assuming that more than one detection error is extremely unlikely.

7.4-1 In Example 7.2, when the sequence $S =$ **101010100000111** was applied to the input of the scrambler in Fig. 7.19a, the output T was found to be **101110001101001**. Verify that when this sequence T is applied to the input of the descrambler in Fig. 7.19b, the output is the original sequence, $S =$ **101010100000111**.

7.4-2 A scrambler is shown in Fig. P7.4-2. Design the corresponding descrambler. If a sequence $S =$ **101010100000111** is applied to the input of this scrambler, determine the output sequence T. Verify that if this T is applied to the input of the descrambler, the output is the sequence S.

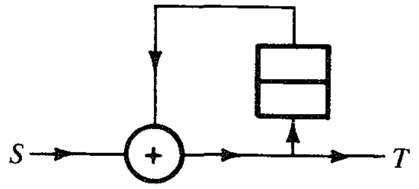


Figure P7.4-2

7.4-3 Repeat Prob. 7.4-2 for the scrambler shown in Fig. P7.4-3.

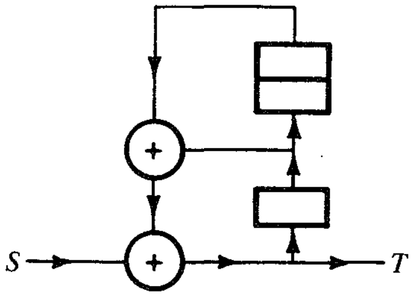


Figure P7.4-3

7.5-1 In a certain binary communication system that uses Nyquist criterion pulses, a received pulse $p_r(t)$ (see Fig. 7.21a) has the following values at the sampling instants:

$$p_r(0) = 1$$

$$p_r(T_b) = 0.1 \qquad p_r(-T_b) = 0.3$$

$$p_r(2T_b) = -0.02 \qquad p_r(-2T_b) = -0.07$$

Determine the tap settings of a three-tap equalizer.

7.6-1 For binary signaling with the received peak amplitude $A_p = 0.0015$ determine the detection error probability if the channel noise is gaussian with rms value 0.0003.

(a) Assuming zero ISI, what is the error probability for: **(i)** polar signaling; **(ii)** on-off signaling; **(iii)** bipolar signaling?

(b) Determine the received power in each of these three cases assuming half-width rectangular pulses.

(c) In order to achieve $P(\epsilon)$ identical to that in the polar case, what must be the received power for the on-off and the bipolar cases?

7.6-2 Half-width rectangular pulses are transmitted at a rate of 10 kHz using an on-off scheme. The detection-error probability is required to be less than 10^{-6}. The rms value of the channel noise at the receiver input is 1 mV. The signal attenuation over the channel (from the transmitter to the receiver) is 30 dB. Determine the minimum signal power that must be transmitted. For simplicity, assume that the pulse shape remains unchanged during the transmission. *Hint*: Knowledge of $P(\epsilon)$ yields the minimum A_p required at the receiver. From this, compute the power at the receiver for half-width on-off pulses.

7.6-3 Repeat Prob. 7.6-2 for polar and bipolar signals.

7.7-1 In multiamplitude scheme with $M = 16$,

(a) Determine the minimum transmission bandwidth required to transmit data at a rate of 12,000 bit/s with zero ISI.

(b) Determine the transmission bandwidth if Nyquist criterion pulses with a roll-off factor $r = 0.2$ are used to transmit data.

7.7-2 An audio signal of bandwidth 4 kHz is sampled at a rate 25% above the Nyquist rate and quantized. The quantization error is not to exceed 0.1% of the signal peak amplitude. The resulting quantized samples are now coded and transmitted by 4-ary pulses.

(a) Determine the minimum number of 4-ary pulses required to encode each sample.

(b) Determine the minimum transmission bandwidth required to transmit this data with zero ISI.

(c) If Nyquist criterion 4-ary pulses with 25% roll-off are used to transmit this data, determine the transmission bandwidth.

7.7-3 Binary data is transmitted over a certain channel at a rate of R_b bit/s. To reduce the transmission bandwidth, it is decided to transmit this data using 8-ary multiamplitude signaling.

(a) By what factor is the bandwidth reduced?

(b) By what factor is the transmitted power increased, assuming the minimum separation between pulse amplitudes to be the same in both cases? *Hint:* Take the pulse amplitudes to be $\pm A/2$, $\pm 3A/2$, $\pm 5A/2$, and $\pm 7A/2$ so that the minimum separation between various amplitude levels is A (the same as in the binary case pulses $\pm A/2$). Assume all the eight levels equally likely.

7.7-4 Consider a case of binary transmission using polar signaling that uses half-width rectangular pulses of amplitudes $A/2$ and $-A/2$. The data rate is R_b bit/s.

(a) What is the minimum transmission bandwidth and the transmitted power?

(b) This data is to be transmitted by M-ary rectangular half-width pulses of amplitudes $\pm A/2$, $\pm 3A/2$, $\pm 5A/2$, ..., $\pm[(M-1)/2]A$. Note that the minimum pulse amplitude separation is A in order to maintain about the same noise immunity. If each of the M-ary pulses is equally likely to occur, show that the transmitted power is

$$P = \frac{(M^2-1)A^2}{24\log_2 M} \approx \frac{M^2A^2}{24\log_2 M}$$

Also determine the transmission bandwidth.

7.7-5 An analog signal of bandwidth 10 kHz is sampled at a rate of 24 kHz, quantized into 256 levels, and coded using M-ary multiamplitude pulses satisfying the Nyquist criterion with a roll-off factor $r = 0.2$. A 30-kHz bandwidth is available to transmit the data. Determine the smallest acceptable value of M.

7.8-1 Figure P7.8-1 shows a binary data transmission scheme. The baseband signal generator uses full-width pulses and polar signaling. The data rate is 1 Mbit/s.

(a) If the modulator generates a PSK signal, what is the bandwidth of the modulated output?

(b) If the modulator generates FSK with the difference $f_{c_1} - f_{c_0} = 100$ kHz (see Fig. 7.29c), determine the modulated signal bandwidth.

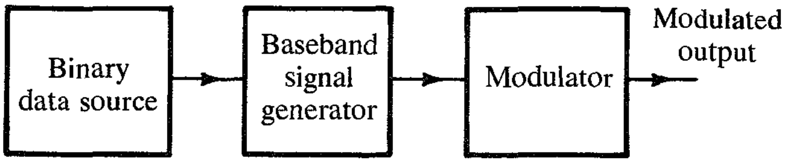


Figure P7.8-1

7.8-2 Repeat Prob. 7.8-1 if instead of full-width pulses, Nyquist criterion pulses with $r = 0.2$ are used.

7.8-3 Repeat Prob. 7.8-1 if a multiamplitude scheme with $M = 4$ (polar signaling with full-width pulses) is used. In FSK [Prob. 7.8-1, part **(b)**], assume that successive amplitude levels are transmitted by frequencies separated by 100 kHz.

7.9-1 In a certain telemetry system, there are four analog signals $m_1(t)$, $m_2(t)$, $m_3(t)$, and $m_4(t)$. The bandwidth of $m_1(t)$ is 3.6 kHz, but those of the remaining signals are 1.4 kHz each. These signals are to be sampled at rates no less than their respective Nyquist rates, and are to be word-by-word multiplexed. This can be achieved by multiplexing the PAM samples of the four signals and then binary coding the multiplexed samples (as in the case of the PCM T1 carrier in Fig. 6.15a). Suggest a suitable multiplexing scheme for this purpose. What is the commutator frequency (in rotations per second)? *Note:* In this case you may have to sample some signal(s) at rates higher than their Nyquist rates.

7.9-2 Repeat Prob. 7.9-1 if there are four signals $m_1(t)$, $m_2(t)$, $m_3(t)$, and $m_4(t)$ with bandwidths 1200 Hz, 700 Hz, 300 Hz, and 200 Hz, respectively. *Hint:* First multiplex m_2, m_3, and m_4 and then multiplex this composite signal with m_1.

7.9-3 A signal $m_1(t)$ is band-limited to 3.6 kHz, and the three other signals, $m_2(t)$, $m_3(t)$, and $m_4(t)$, are band-limited to 1.2 kHz each. These signals are sampled at the Nyquist rate and binary coded using 512 levels ($L = 512$). Suggest a suitable bit-by-bit multiplexing arrangement (as in Fig. 7.33c or d). What is the commutator frequency (in rotations per second), and what is the output bit rate?

8 EMERGING DIGITAL COMMUNICATIONS TECHNOLOGIES

WILLIAM J. JAMESON

The term **digital communications** means different things to different people. It often includes digitally modulated sinusoidal signals (e.g., phase and frequency shift keying). In the context of this section, however, it will refer to electrical or optical signals transmitted in discrete pulses. Such signals typically share a common characteristic: They are transmitted in **frames** or **cells**, which allow synchronization between transmitter and receiver. We shall investigate several types of such framed digital signals, noting that, historically, they arose in the telephone industry. First of all, however, we shall consider the very mature "digital hierarchy" on which the emerging technologies are based.

8.1 THE NORTH AMERICAN HIERARCHY

The first truly digital signal to be used extensively was the pulse-code-modulated (PCM) voice signal. Recall that the typical PCM voice signal consists of a nominal 4000-Hz analog voice signal sampled with 8 bits of quantization, representing amplitude, 8000 times per second. In actual practice, primarily to avoid aliasing, the analog signal is band-limited to 300 to 3300 Hz before sampling. A sampled voice signal, therefore, represents 64 kbit/s, that is, 8000 samples per second × 8 bits per sample. Such a signal is referred to as a **digital signal at level zero (DS0)**. Note that each sample is transmitted in 1/8000 of a second, that is, 125 μs. It is the fundamental building block of digital communications, and we shall see that many signals are based on a 125-μs transmission time.[1]

A single DS0 signal is virtually never transmitted by itself. This is primarily due to the fact that most telephones are analog devices. Hence, in the typical telephone central office, a subscriber's analog line is first terminated in a (300 to 3300-Hz) band-limiting or antialiasing filter which, in turn, is terminated in a codec (coder/decoder) that converts the analog signal into a DS0. For network transmission (e.g., between telephone central offices), 24 of these DS0s are multiplexed into a DS1 (digital signal at level 1), more commonly (but not always correctly) referred to as a T1 signal. It is at this point that the multiplexed signal is packaged into a frame. There are two framing formats which are currently used, at least in the United States. One is the D4 frame and the other the extended superframe (ESF).

8.1.1 D4 Framing (SF)

The D4 frame packages 24 bytes, each representing 8 quantization bits, into a 193-bit frame. The 193 bits represent 24 channels × 8 bits = 192 bits plus a **framing bit**; and 12 consecutive frames comprise a **superframe (SF)**, which is transmitted in 1.5 μs. Hence, 12 frames × 193 bits ÷ 1.5 μs yields 1,544,000 bit/s. A DS1 or T1 is, therefore, a 1.544-Mbit/s digital signal.

What is the purpose of the framing bits? Unfortunately, the clocks at the transmit and receive ends of a DS1 transmission may be slightly out of common synchronization. Hence, the framing bits define a fixed pattern, which allows the receiver to identify each superframe and, hence, each of the 24 digital streams for proper demultiplexing. Suppose there were no framing bits? Figure 8.1 illustrates the possible problem that could result.

In Fig. 8.1, τ is the time required for two T1 transmitted/received bits to become out of synchronization (due to clock differences between the transmit and receive ends of the transmission) by 1 bit. In a typical (stratum 4) telephone central office or private-user telephone switch Private Branch eXchange (PBX) the required synchronization accuracy is $\pm 3.2383 \times 10^{-5}$, or ±50 bits in 1.544 Mbits. Without framing to provide a means of synchronization between transmitter (TX) and receiver (RX), time slots will overlap, causing distortion and, eventually, the interchange of two time slots. In unframed T1 (e.g., for high-speed data transfer) other methods are used to maintain TX/RX synchronization.

The D4 framing (bit) sequence (the sequence of 12 framing bits in a D4 superframe) is **100011011100**, as pictured in Fig. 8.2a. The framing sequence provides a fixed reference to the telephone system network synchronization.

In DS1 transmission one needs to provide information related to the establishment of a calling path (the "setup" and "knockdown" of a call, for example) and other system control

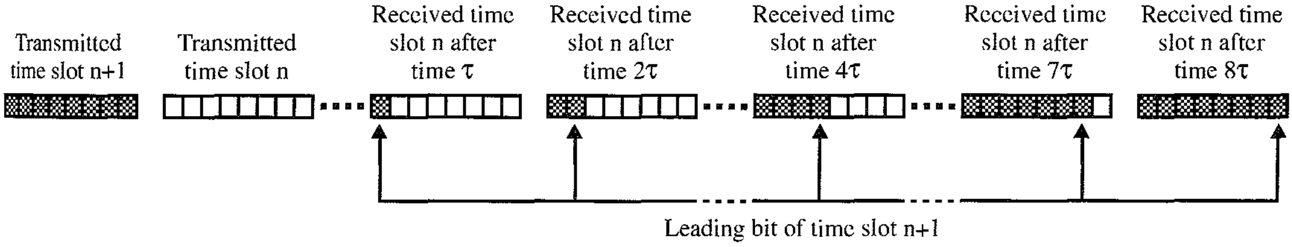

Transmitted time slot n+1
Transmitted time slot n
Received time slot n after time τ
Received time slot n after time 2τ
Received time slot n after time 4τ
Received time slot n after time 7τ
Received time slot n after time 8τ
Leading bit of time slot n+1

Figure 8.1 Consequences of free-running (unframed) T1 in voice transmission.

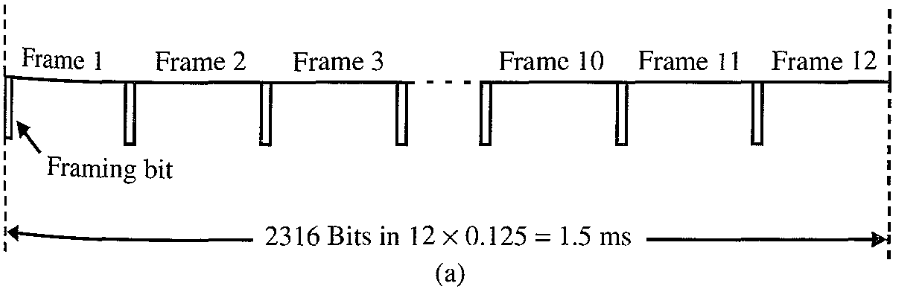

Frame 1
Frame 2
Frame 3
Frame 10
Frame 11
Frame 12
Framing bit
2316 Bits in 12 × 0.125 = 1.5 ms
(a)

Figure 8.2a Framing bit sequence.

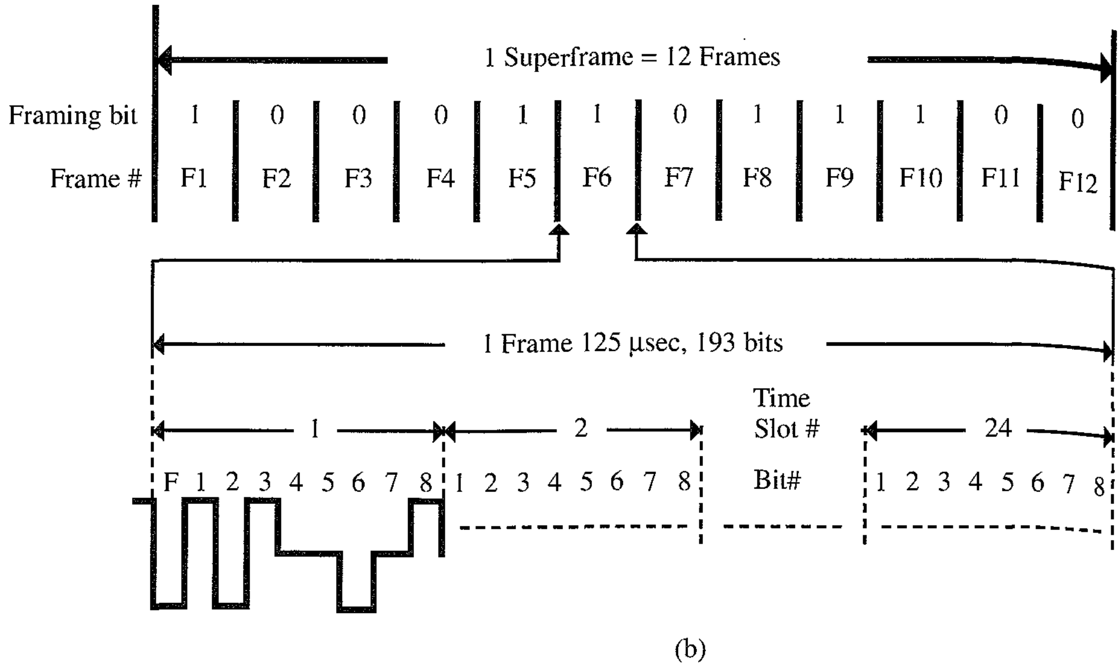

1 Superframe = 12 Frames
Framing bit
1
0
0
0
1
1
0
1
1
1
0
0
Frame #
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
F11
F12
1 Frame 125 μsec, 193 bits
1
2
Time Slot #
24
F 1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
Bit#
1 2 3 4 5 6 7 8
(b)

Figure 8.2b D4 framing.

information. This is done by means of **robbed bit signaling**. In order to do this in D4 framing, the least significant (eighth) bit of each byte in each sixth frame is "robbed" as a signaling bit, as shown in Fig. 8.3.

Historically, digital signaling was developed before the availability of modern digital switches that could accommodate digital signals. As a result, a device called a **channel bank** was developed, which performed several basic functions (Fig. 8.4):

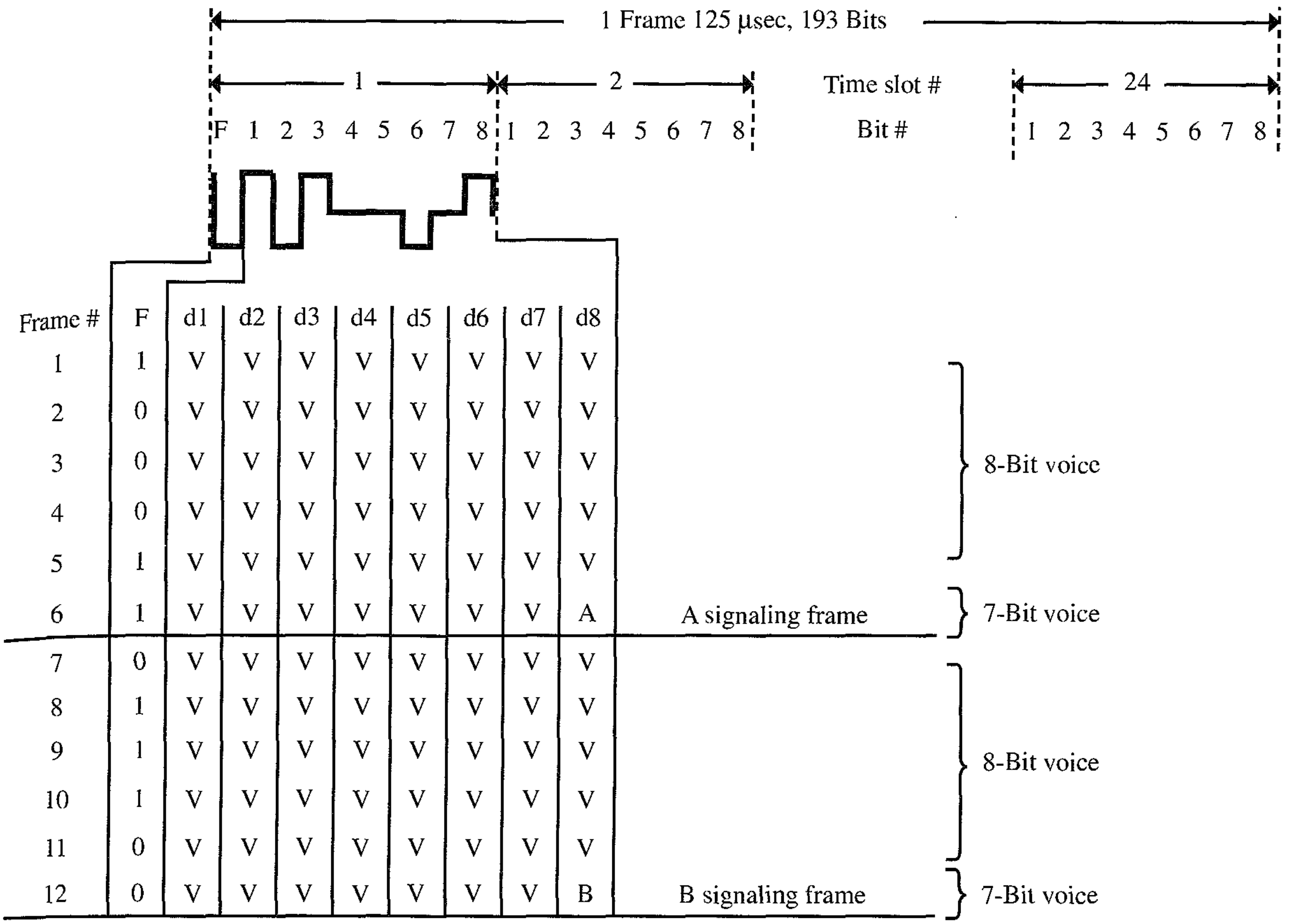
- Band-limiting the signal to approximately 300 to 3300 Hz to avoid aliasing
- Companding the signal to reduce distortion of low-level signals due to quantization
- A/D and D/A conversion
- Multiplexing the individual DS0 signals

Note that the analog lines may be voice or digitally modulated analog data signals at rates of up to 56 kbit/s.

Channel banks were placed at the network side of telephone switches so that more reliable digital transmission could take place between analog switches. Many are still used today, particularly in rural areas that have not yet converted to digital central office switches. Modern channel banks provide D4 or ESF framing.

8.1.2 ESF Framing

As DS1 service expanded to carry many types of customer payloads, it became evident that repetitive patterns in such payloads would often counterfeit the D4 framing pattern shown in Fig. 8.2b. This would either cause the DS1 receiver to frame on the wrong position and



Frame #	F	d1	d2	d3	d4	d5	d6	d7	d8	
1	1	V	V	V	V	V	V	V	V	
2	0	V	V	V	V	V	V	V	V	
3	0	V	V	V	V	V	V	V	V	
4	0	V	V	V	V	V	V	V	V	
5	1	V	V	V	V	V	V	V	V	
6	1	V	V	V	V	V	V	V	A	A signaling frame
7	0	V	V	V	V	V	V	V	V	
8	1	V	V	V	V	V	V	V	V	
9	1	V	V	V	V	V	V	V	V	
10	1	V	V	V	V	V	V	V	V	
11	0	V	V	V	V	V	V	V	V	
12	0	V	V	V	V	V	V	V	B	B signaling frame

V-Voice code bit, A-primary signaling bit, B-secondary signaling bit

Figure 8.3 Robbed bit signaling.

produce gibberish as output, or it would prevent framing entirely in channel banks equipped to inhibit the declaration of an "in frame" condition when two or more valid framing patterns were present.

In order to prevent false or inhibited framing, a new framing format was developed, called the **extended superframe (ESF)**. ESF divides the 8-kbit/s framing channel into three separate streams. In addition, ESF uses a superframe of 24 frames in length, rather than the 12 frames of D4 (or SF). To differentiate it from ESF, the D4 format is known in standards documents as superframe format (SF).

Two kbit/s of the framing, or **overhead**, are allocated to framing the DS1 receiver. Six of the 24 overhead bits per ESF are thus assigned to framing. Two kbit/s of overhead are assigned to error checking in the form of a **cyclic redundancy check (CRC)**. The 6 bits per ESF made available for this function are assigned as a CRC-6 error-checking sequence, the operation of which is defined in telecommunications standard ANSI T1.403. When framing, a receiver can find a candidate alignment and then check the CRC-6 bits for a match between the locally generated remainder and the remainder received from the distant terminal. If the two CRC-6 remainders do not match, the framing alignment is rejected as a probable counterfeit. The

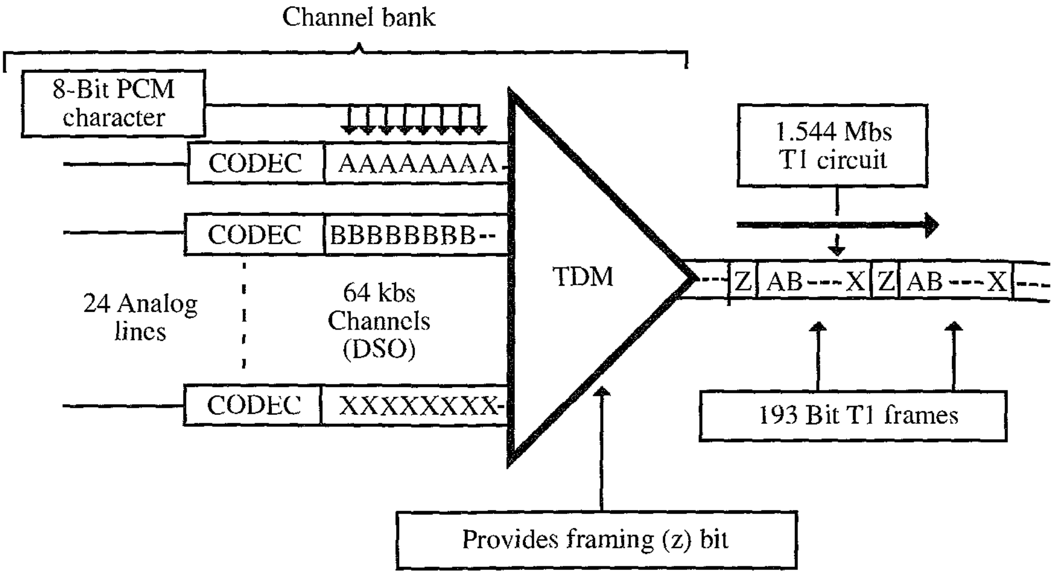


Figure 8.4 Channel bank.

CRC-6 sequence is also valuable for detecting errors while the DS1 is in service and makes possible more comprehensive performance monitoring with ESF than is possible with D4. Note that, although error correcting can be performed by cyclic redundancy checks, only 6 CRC bits are provided to check 4608 payload bits in the ESF format. This is sufficient for reliable error checking, but not adequate for error correction. The remaining 4-kbit/s overhead capacity is assigned as a **data link (DL)**, which is used to carry performance monitoring information, alarms, and commands to equipment from one location to another.

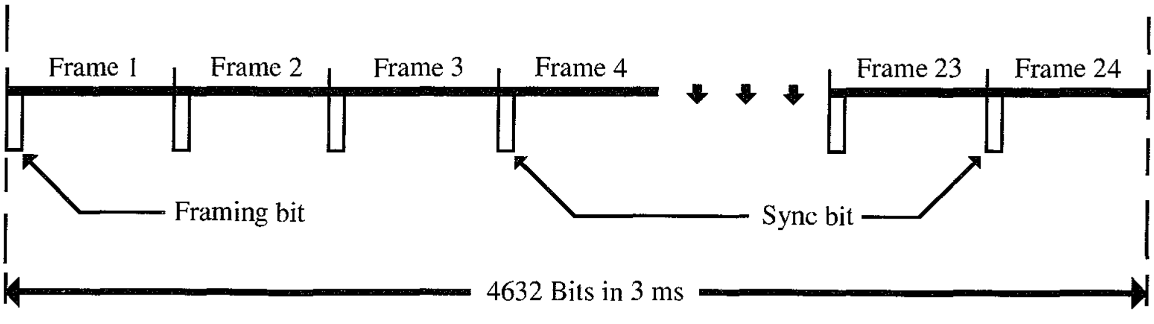
There are four signaling options in ESF, but only the T-Option uses the 8th bit in each sixth frame for information coding. Fig. 8.5c illustrates the various signaling options. Note that Figs. 8.5a and b illustrate only the T-Option ESF framing sequence.

Since there are 24 frames per ESF, there are signaling bits carried in the "robbed" eighth bits of the channels in frames 6, 12, 18, and 24. Note that there are four different signaling bits per channel aligned with the framing pattern, and these are labeled A, B, C, and D in lieu of the A and B of SF. It is possible, therefore, to transport 16-state signaling through an ESF DS1 while an SF-formatted system can handle, as a maximum, four-state signaling. In most cases, channels in both formats are used to provide simpler two-state signaling, but there are special-services applications for the more comprehensive signaling in each case.

Most equipment that supports either D4 or ESF framing also makes it possible to eliminate the writeover of the eighth bit by signaling. This, in combination with a bit-sequence-independent line code such as B8ZS, makes it possible to transport 64 kbit/s of customer data over each channel of either the SF or the ESF system (Fig. 8.5). The cyclic redundancy check provides a means to detect bit errors (see, for example, Stallings[2]).

8.1.3 Regeneration

The viability of regenerative repeaters is the greatest advantage of digital communication over the analog. If the sampling rate exceeds the Nyquist rate for the band-limited signal sufficiently, a received digital signal is normally superior to a corresponding analog signal due to the use



(a)

Figure 8.5a ESF framing bit sequence.

of regenerative repeaters, as explained in Sec. 7.5. In the analog signal, noise is added in transmission, in switching systems, and in repeaters, among others. In a digital signal, however, the added noise is removed by regenerating the original signal instead of simply repeating it, along with the noise. The **regenerator** is a device that performs this function, as shown in Fig. 8.6 (see Sec. 7.5).

8.1.4 DS1/T1 Transmission

There are several format and transmission requirements of a DS1/T1, as described in Fig. 8.7. For voice transmission, the DSX1 interface is required. For data transmission, there are no restrictions on format.

The DSX1 interface defines the basic DS1 transmission requirements, including the physical connections. The voltages are associated with the **line code**, the bit transmission format. The DS1 transmission uses the **alternate mark inversion (AMI)** line code, sometimes referred to as **bipolar, return to zero (BPRZ)**, as pictured in Fig. 8.8.

The AMI line code is used for two purposes: to eliminate a dc voltage on the transmission line and to extract timing information for transmission path synchronization. Note that extraction of timing information is based on maintaining a charge on a tuned circuit at each end of the transmission path. The tuned circuit is charged by the marks. Hence, if there are no marks transmitted over a short period of time, the timing information is lost. Hence, the DSX1 requires a mark every 8 bits, or at least three marks in 23 bits (DSX1 average bit density, or **1**'s density, specification). Moreover, there cannot be 16 consecutive transmitted zeros (DSX1 sixteen **0**'s specification). For data transmission, the consecutive density problems of zeros and **1**'s typically do not exist. Synchronization is maintained by error detection/correction protocols.

EXAMPLE 8.1 Following are examples of a **1**'s density and a 16 **0**'s violation.

(a) **1**'s density violation:

$\mathbf{11}\underline{\mathbf{010001000000001000000000}}\mathbf{111}$ (three **1**'s in 24 bits)

(b) 16 **0**'s violation:

$\mathbf{111011}\underline{\mathbf{0000000000000000}}\mathbf{11011}$ (16 consecutive **0**'s)

ESF Frame Number	F-Bits			
	ESF Bit Number	Assignments		
		FPS	FDL	CRC
1	0	-	M	-
2	193	-	-	CB1
3	386	-	M	-
4	579	0	-	-
5	772	-	M	-
6	965	-	-	CB2
7	1158	-	M	-
8	1351	0	-	-
9	1544	-	M	-
10	1737	-	-	CB3
11	1930	-	M	-
12	2123	1	-	-
13	2316	-	M	-
14	2509	-	-	CB4
15	2702	-	M	-
16	2895	0	-	-
17	3088	-	M	-
18	3281	-	-	CB5
19	3474	-	M	-
20	3667	1	-	-
21	3860	-	M	-
22	4035	-	-	CB6
23	4246	-	M	-
24	4439	1	-	-

Figure 8.5b ESF Bit Assignments

FPS - Framing Pattern Sequence

FDL - 4 kbit/s Facility Data Link Message

CRC - Cyclic Redundancy Check

CBx - Block Check Field Bit # x

(b)

8.1.5 B8ZS Line Code

A typical digital signal may have several long strings of **0**'s. Timing information in the network is derived from the data signal; **0**'s provide no timing information. Only **1**'s provide the required timing information. At regenerators and at the receiving end, the clock signal is derived from

			ESF Signaling Options			
Frame number	Information coding bits	Signaling bit #	T	2 State	4 State	16 State
1	1 - 8					
2	1 - 8					
3	1 - 8					
4	1 - 8					
5	1 - 8					
6	1 - 7	8	-	A	A	A
7	1 - 8					
8	1 - 8					
9	1 - 8					
10	1 - 8					
11	1 - 8					
12	1 - 7	8	-	A	B	B
13	1 - 8					
14	1 - 8					
15	1 - 8					
16	1 - 8					
17	1 - 8					
18	1 - 7	8	-	A	A	C
19	1 - 8					
20	1 - 8					
21	1 - 8					
22	1 - 8					
23	1 - 8					
24	1 - 7	8	-	A	B	D

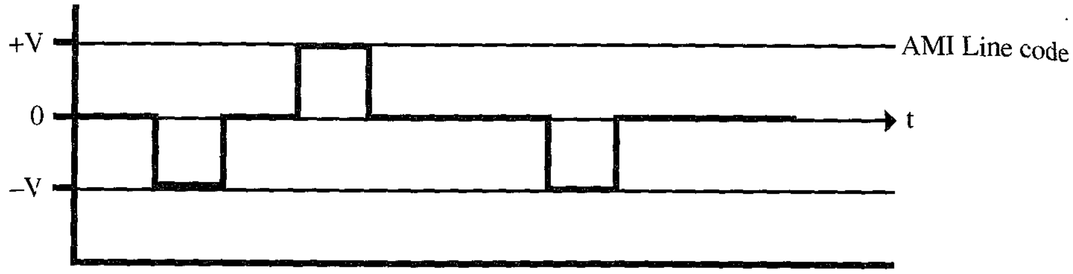
Options: T - Transparent; 8th bit used for information coding
2 State - Provides one 1333 bit/s signaling channel (A)
4 State - Provides two 667 bit/s signaling channels (A & B)
16 State - Provides four 333 bit/s signaling channels (A, B, C & D)

(c)

Figure 8.5c ESF signaling options

a tuned circuit driven by the **1**'s pulses. Without sufficient **1**'s, the circuit damps down to the point at which it can no longer provide timing information and the line loses synchronization. Hence (in the U.S. T1 standard), no more than 15 **0**'s can be sent in succession; there must be a **1**'s density of at least three **1**'s each 23 bits to ensure proper timing. Idle circuit **1**'s are inserted by the DSU to avoid loss of sync.

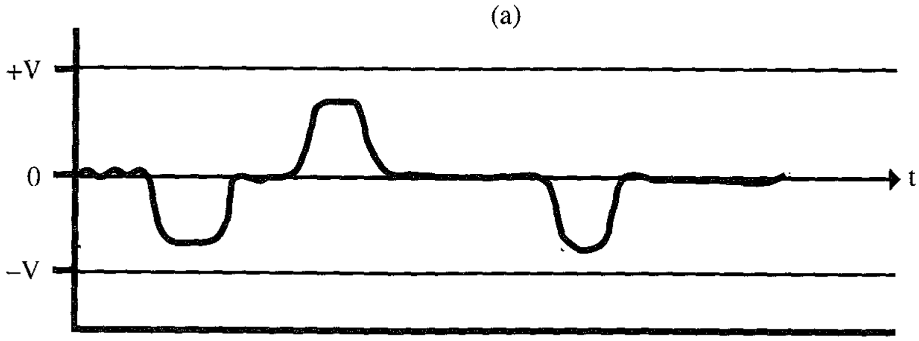
A clever line code was developed which corrects these problems automatically in a manner that is easily accommodated by the channel service unit (CSU)/digital service unit (DSU). It is called the **binary 8-zeros suppression (B8ZS)** line code (see Sec. 7.2). Whenever eight successive **0**'s are detected, the implementation of this line code produces the automatic insertion of a special 8-bit sequence containing an intentional bipolar violation that can be detected and corrected easily by the CSU/DSU. The violations distinguish a byte substituted for all **0**'s from a normal byte containing legal **1**'s. B8ZS is used to transmit 64-kbit/s clear DS0 data channels. (AMI requires the 3 in 23 **1**'s density; hence, it cannot be used for data



+V
0
–V
AMI Line code
t

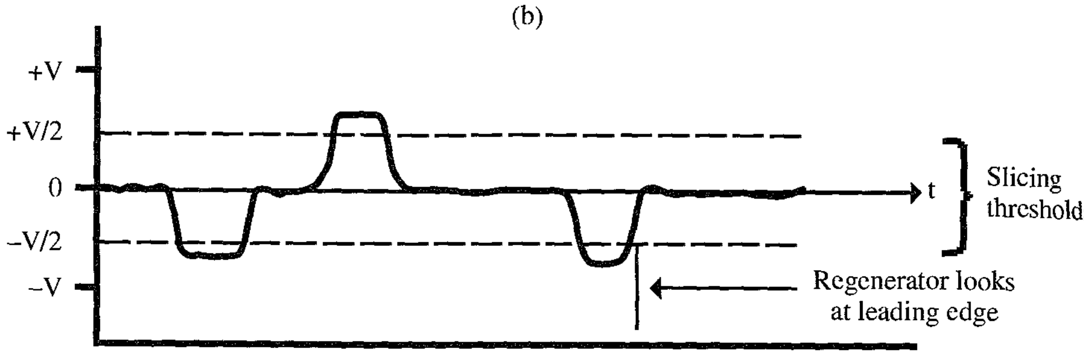
Original signal

(a)



Signal plus noise added in transmission

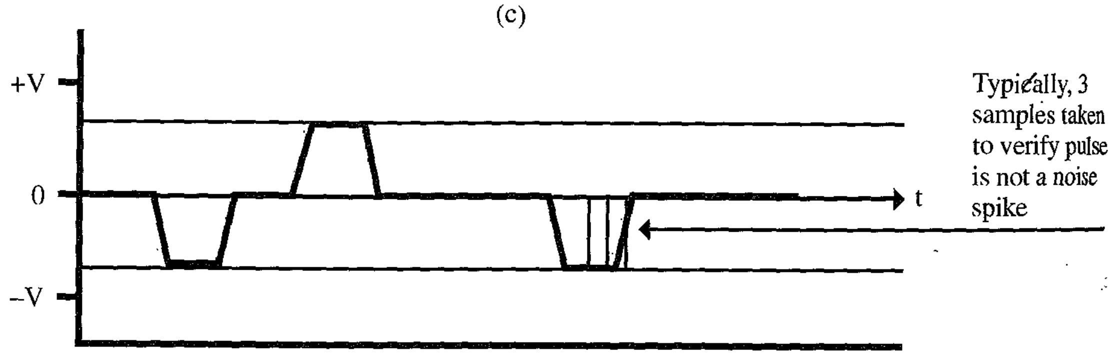
(b)



+V
+V/2
0
-V/2
–V
t
Slicing thresholds
Regenerator looks at leading edge

Noise slicing

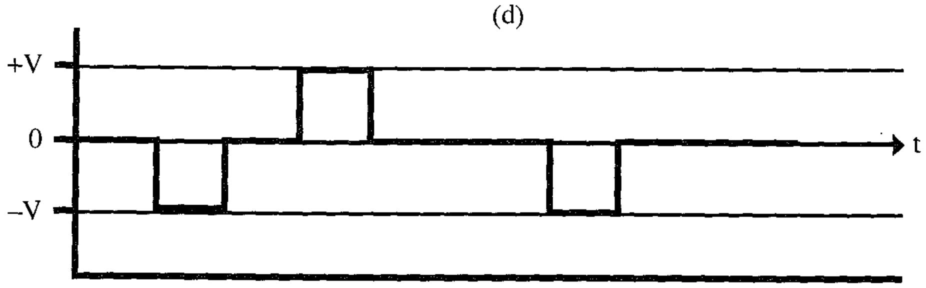
(c)



+V
0
–V
t
Typically, 3 samples taken to verify pulse is not a noise spike

Sliced & sampled noisy signal

(d)



Regenerated signal

(e)

Figure 8.6 Regeneration.

DS1
(Digital Signal, Level 1 - 1.544 Mb/s)

DSX1
(Digital Signal Cross Connect, Level 1)
- Connector
- Pin placement
- Voltages
- 16 Zeros
- Average bit density

Data
- Any data format

Figure 8.7 Transmission interfaces.

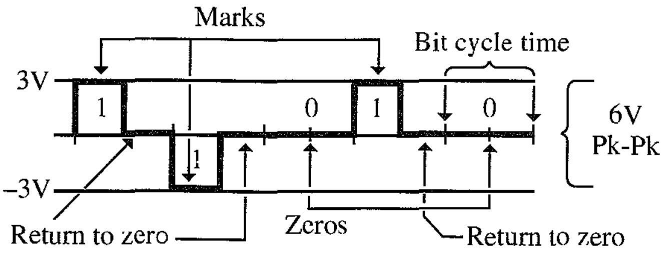

Marks
Bit cycle time
3V
1
1
0
1
0
6V
Pk-Pk
−3V
Return to zero
Zeros
Return to zero

Figure 8.8 AMI line code.

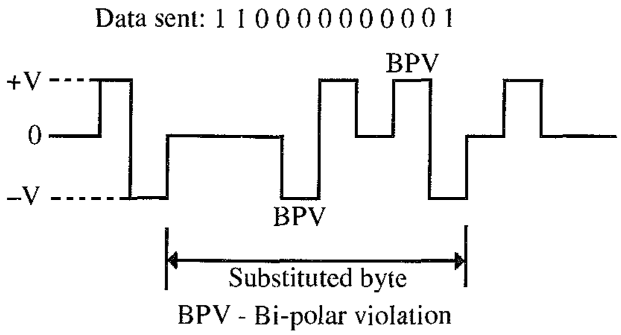

Data sent: 1 1 0 0 0 0 0 0 0 0 0 1
BPV
+V
0
−V
BPV
Substituted byte
BPV - Bi-polar violation

Figure 8.9 B8ZS line code.

transmission at 64 kbit/s because the insertion of **1**'s at idle tone disturbs the data stream.) Note that B8ZS allows no more than eight consecutive **0**'s, and the bipolar violation pattern uniquely identifies the eight **0**'s. Figure 8.9 illustrates the procedure. Also note that the voltage levels in the "violating byte" of this scheme average to zero, so that no dc is allowed to build up on the circuit.

EXAMPLE 8.2 The signal **11010000000001** is received by the DSU in a T1 data stream which uses a B8ZS format. Draw the output of the DSU for this signal. Identify the 8-bit stream that is replaced by the DSU.

8.1.6 DSU and CSU

Clearly, there must be some mechanism to detect and, hopefully, correct errors in transmission due to **bipolar violations** (two consecutive marks sent with the same polarity), 16 or more consecutive spaces (**0**'s), or fewer than three marks in 23 bits. There are two devices that perform these functions: the **digital service unit (DSU)** and the **channel service unit (CSU)** (Fig. 8.10).

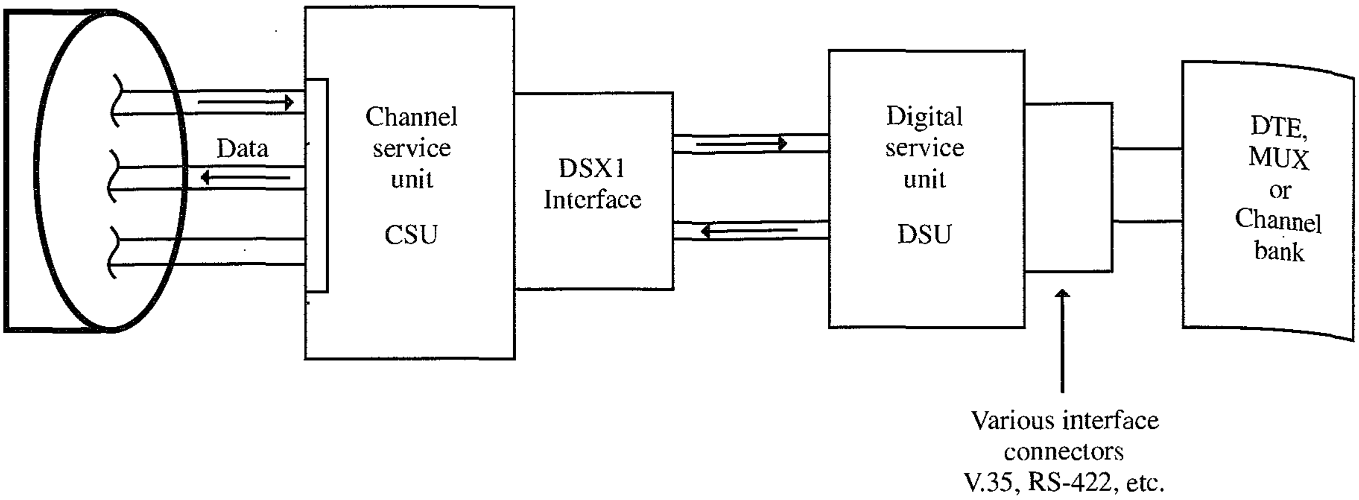


Figure 8.10 CSU and DSU.

The CSU contains the following elements:

- The repeater (regenerator) section, which sends and receives data and regenerates the signal.
- The signal monitor section, which performs the 16 **0**'s and average bit density (three marks in 23 bits) checks.
- A remote test section, which conducts a loop-back path and provides a customer indication of the test.

The DSU performs the following functions:

- It converts standard synchronous (data communication equipment and data terminal equipment) interface to DSX1 interface.
- It prevents bipolar violations (this may involve shutting down the circuit).
- It maintains the 16 **0**'s and average bit density rules.
- It provides unframed or framed network compatibility.

Note that, functionally, the DSU and the CSU are often combined into one unit.

8.1.7 High-Bit-Rate Digital Subscriber Line (HDSL)

The high-bit-rate digital subscriber line (HDSL) is often lumped in with other technologies that increase the capacity of the copper telephone cable pairs, such as ADSL and VDSL (see Sec. 8.5.3). HDSL is a bit different, however, in that its sole purpose is to transport a DS1 signal through cable pairs that would not support transmission using the AMI or B8ZS line codes of the T1 line. HDSL is used in lieu of T1 lines in a significant fraction of the new DS1 services, which are being installed today. It might be said that HDSL is simply a better line code for DS1 transmission, but it involves more than that. There are various HDSL formats, which use different numbers of cable pairs to transport the signal, but the one described here is used almost universally in North American installations.

HDSL splits the 1.536-Mbit/s DS1 payload into two bit streams of 768 kbit/s each. To each is added the full 8-kbit/s overhead of the DS1 signal. The overhead consists of either the D4 frame bits or the ESF frame bits, CRC-6 bits, and data link bits. To each bit stream is added an additional 8-kbit/s overhead for framing and administering the HDSL system, bringing the bit rate of each stream up to 784 kbit/s.

Each 784-kbit/s signal is encoded using the same 2B1Q line code that is used for ISDN BRI, and which is discussed in detail elsewhere in this chapter. Encoding the HDSL signal using 2B1Q results in a baud rate of 392 K baud and a power density spectrum that peaks at somewhat below 200 kHz. The maximum power density spectrum with either AMI or B8ZS is about 772 kHz. HDSL thus brings the power density spectrum down in frequency, reducing the effects of several impairments such as near-end and far-end cross talk, bridged taps, external interference, and cable attenuation.

Note that HDSL requires two cable pairs to transmit the same payload that a T1 line would carry on one pair. In order to avoid the need for four pairs to provide a complete two-way transmission path, HDSL "reuses" the same pairs for transmission in the opposite direction. Interference between transmitter and receiver at each end of such a circuit is prevented by using digital echo cancellation techniques.

8.1.8 Digital Hierarchy

We have discussed the DS0 and DS1 formats. The DS0 is multiplexed into a DS1 with the framing bits to provide TX/RX synchronization. Most often, for network transmission, one wishes to multiplex DS1s into higher rate signals. Hence, there are several higher level digital signals that may be used. These form the **digital hierarchy**, as noted in Table 8.1.[3]

The advent of fiber-optic transmission systems has reduced the importance of the DS4 and we shall see that the optical transmission formats have superseded the DS4. The DS3, however, is used extensively for transmission of multiplexed voice signals, digital television signals, and high-speed data transfer.[1]

The DS2 and DS1C are used mostly by users who require a higher bandwidth than that provided by DS1. A great many small digital microwave systems, for example, use the DS2 transmission format. A typical hierarchical structure is shown in Figure 8.11. Table 8.2 defines the multiplexing relationships.

Table 8.1
Digital Transmission Formats—Digital Hierarchy

Level	Bit rate	# DS0 circuits	Control overhead
DS0	64 kbit/s	1	0 bits
DS1	1.544 Mbit/s	24	8000 bit
DS1C	3.152 Mbit/s (T1c)	48 (2 DS1s)	80000 bit
DS2	6.312 Mbit/s (T2)	96 (4 DS1s)	168,000 bit
DS3	44.736 Mbit/s (T3)	672 (28 DS1s)	1.728 Mbit
DS4	274.176 Mbit/s (T4)	4032 (422 DS1s)	16.128 Mbit

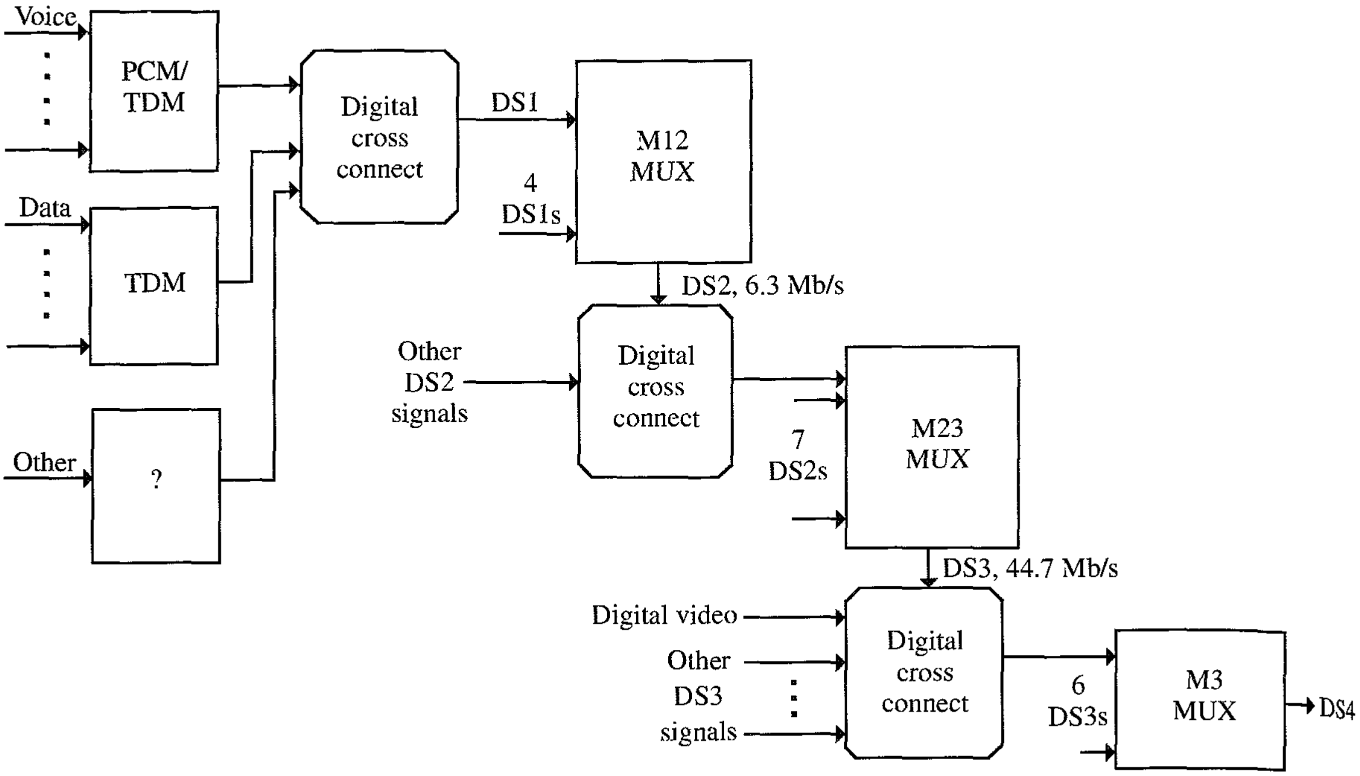

Voice
PCM/ TDM
Data
TDM
Other
?
Digital cross connect
DS1
4 DS1s
M12 MUX
DS2, 6.3 Mb/s
Other DS2 signals
Digital cross connect
7 DS2s
M23 MUX
DS3, 44.7 Mb/s
Digital video
Other DS3 signals
Digital cross connect
6 DS3s
M3 MUX
DS4

Figure 8.11 Typical digital network structure.

Table 8.2
Digital Multiplexing in Figure 8.11

Multiplexer symbol	Multiplexes
M12	4 DS1s → 1 DS2
M13	28 DS1s → 1 DS3
M23	7 DS2s → 1 DS3
M3	6 DS3s → 1 DS4

Note that the number of control bits increases rapidly as one progresses along the digital hierarchy. Recall that we observed that two network nodes might have slightly different internal clock rates. If one were to multiplex 28 DS1s into a DS3, there could be 28 different clock rates that the M13 multiplexer would have to accommodate. This accommodation is accomplished by means of control bits called **stuff bits**. The multiplexed DS1 signals all must have the same transmission rate within the DS3. This is accomplished by adding stuff bits to the slower DS1 bit streams so that the same number of bits per second is transmitted for each DS1 in the multiplexed data stream. Fortunately, a DS1 clock must perform within well-defined tolerances, which can be accommodated by the stuff bits.

We should also observe that a digital cross connect is not a multiplexer in the strict sense. It simply cross connects several digital signals into a single data stream (see, for example, Bellamy[3]).

8.1.9 Network Synchronization

Prior to the divestiture in 1984, network synchronization was maintained by AT&T from their master control center in a suburb of St. Louis. Each of several levels in the AT&T network hierarchy maintained a submaster clock synchronized to the next higher level. After the divestiture, however, some of the Bell operating companies (BOCs) decided that it was in their best interest to maintain their own synchronization by means of **plesiochronous** (from the Greek word *plesios*, meaning near or nearly) networks. These networks typically operate from cesium beam clocks, which are accurate to ± 3 parts in 10^{12}. In the BOC networks, for example, network synchronization was originally maintained by a timing signal transmitted by LORAN-C from a master cesium clock.

Network synchronization is increasingly using the global positioning system (GPS) to derive universal-time-coordinated (UTC) information. UTC allows precise frequency/timing comparison for shorter time intervals than conventional HF (LORAN-C) signals. Note that a bit error due to synchronization drift can occur every $10^{12}/(2 \times 3)$, or 1.667×10^{11}, bits. (The factor of 2 in the denominator arises from the fact that both the transmit and the receive end can be out of synchronization.) Each local access transport area (LATA) has either a GPS receiver, a LORAN-C receiver, or a rubidium clock.

EXAMPLE 8.3 Two plesiochronous digital networks, A and B, utilize cesium beam clocks accurate to 3 parts in 10^{13}. The networks are operated by independent long-distance companies and are synchronized to each other by means of a UTC signal. If a company leases a T1 line with ESF framing which is terminated at one end in network A and at the other end in network B, how often, in hours to the nearest hour, must the networks be resync'd to each other to avoid a framing bit error in the customer's T1 signal in the worst case (both networks out of synchronization in the opposite time direction)? *Note:* A framing bit error can occur when the two networks are out of synchronization by $^{3}1/5$ of a T1 "bit time" (number of seconds per bit) due to the way in which the digital signal is sliced, sampled, and regenerated.

Since both clocks can be out of synchronization by as much as 6 parts (bits) in 10^{13}, we have

$$\begin{aligned}\frac{6 \text{ timing error bits}}{10^{13} \text{ bits}} \times 1,544,000 \text{ bit/s} &= 9.26 \times 10^{-7} \text{ timing error bits per second}\\ &= 1,079,447.3 \text{ seconds per timing error bit}\\ &= 299.85 \text{ hours per timing error bit}\end{aligned}$$

Since a synchronization error can occur whenever the network is out of synchronization by $^{3}1/5$ bit, the time between resync'ing is 60 hours, to the nearest hour.

8.1.10 Classification of Digital Signals

Digital signals are typically classified as narrow-band, wide-band, or broadband. We shall use the classification scheme* of Figure 8.12 to differentiate between digital signals.

---

* Sands, Douglas, Lucent Technologies Bell Labs, Personal Communication.

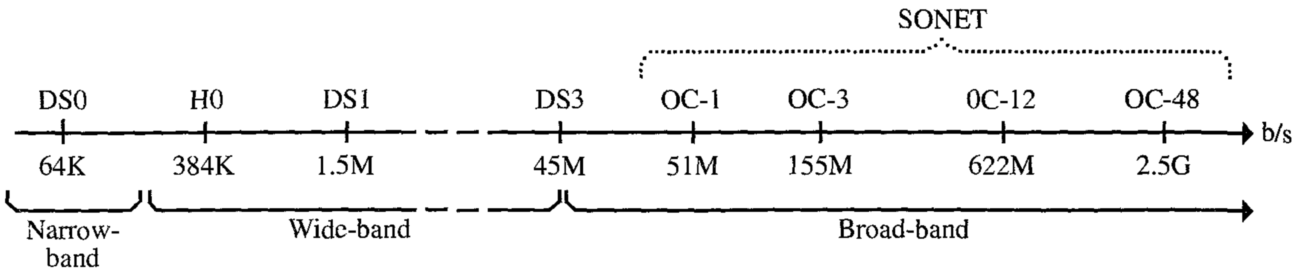


Figure 8.12 Classification of digital signals.

8.2 DIGITAL SERVICES

Classification of Circuits

We shall be concerned with two types of circuits: dedicated and switched.

1. *Dedicated circuit*: A fixed ("permanent") communication circuit between (among) two (or more) locations. It is available at all times without dialing or switching and sometimes is called an "engineered" or "provisioned" circuit. *Example*: A private T1 network.
2. *Switched circuit*: A continuous, fixed-path communication circuit set up in the network. It is available to the user as long as the user wants it and is willing to pay for its use (i.e., as long as the switch is "closed"), and is provided via one or more network switches. *Example*: A telephone call. There are two types of switched circuits:
   - *Circuit switched*: A fixed circuit is set up end to end and maintained during the period of use.
   - *Packet switched*: Information is transmitted in discrete (data/digital) packets, each of which may be transmitted by any of several available different routes. It requires routing control and packet assembly and disassembly. *Note*: Packetized data, particularly in fixed-length packets, may be transmitted on switched as well as on dedicated circuits (e.g., Frame Relay).

The relative advantages and disadvantages of dedicated vs. switched circuits are listed in Table 8.3.

Table 8.3
Relative Advantages of Switched vs. Dedicated Circuits

Dedicated circuit	Switched circuit
Paid for all of the time	Paid for only while switch is closed
High cost of implementation	Low cost of implementation
Always available	Available if network is not loaded to capacity (possibility of busy signals)
Cheap per unit of time if used extensively; expensive per unit of time if used little or only moderately	Cheap if used occasionally or moderately; expensive if used extensively

8.2.1 ISDN

ISDN is an acronym for **integrated services digital network**. It is a switched rather than dedicated digital service, and therein lies its advantage. Traditionally, digital services at, for example, T1 rates (1.544 Mbit/s) required a special, dedicated line and, therefore, were priced accordingly. The cost of a dedicated T1 line was prohibitive unless the subscriber used it extensively. Hence, digital communication was not available to a more casual user. Moreover, the transmission of digital signals via a modem is probably limited to a throughput of 9600 bit/s, even with a 28.8-kbit/s modem. Typically, higher modem throughput may require a special (dedicated) "conditioned" line.

Starting in the late 1970s, the new ISDN switched service was developed, primarily by the T1D1 subcommittee of the Exchange Carriers Association in the United States and by the CCITT (now ITU-T for International Telecommunications Union—Telephony) Study Group XVIII. The activities of these two groups were coordinated and resulted in compatible international ISDN standards (see, for example, Helgert[4]).

The two fundamental structures of ISDN are called **basic rate interface (BRI)** and **primary rate interface (PRI)**. Each is based on multiple 64-kbit/s DS0 **bearer** or **B channels** and a **data** or **D channel**. In the BRI structure, there are two B channels and a 16-kbit/s D channel. In the PRI structure, there are 23 B channels and one 64-kbit/s D channel for a total of 1.536 Mbit/s. Hence, BRI is often called $2B + D$ and PRI is referred to as $23B + D$.

In addition to the B and D channels, there are H channels for situations that require higher data rates. The H0 channels are transmitted at 384 kbit/s, which is a submultiple of the 1.536 Mbit/s of the North American standard PRI and the European CEPT/ITU-T 1.920 Mbit/s versions. The H0 rate is the rate of choice in many interactive video systems, for example.

ISDN Synchronous Channel Structures

B Channel: The B channel operates at a synchronous rate of 64 kbit/s, full duplex. It poses no restriction on the binary representation of the data it carries. It is used for:

- Digital voice encoded at 64 kbit/s, PCM
- Synchronous data streams at 600,1200, 2400,4800, or 9600 bit/s or 48 or 64kbit/s; lower rate data streams may be multiplexed
- Digital voice encoded at less than 64 kbit/s, either alone or multiplexed with other data or voice (e.g., ADPCM, CVSD)

In the latter two cases the aggregate data rate must be 64 kbit/s.

D Channel: The D channel is one of two types, synchronous at 16 kbit/s or at 64 kbit/s, full duplex. Its primary function is to carry signaling information for the control of circuit switched connections involving one or more B channels. When it is not carrying such information it may be used for:

- Transport of user signaling information

- Low-bit-rate packet data
- Telemetry

All of these alternate uses of the D channel use statistical multiplexing and priority access for call control signals to avoid contention.

H Channel (see Table 8.4): H channels are used to carry user information at data rates in excess of 64 kbit/s. Specific applications include:

- High-resolution digital video/audio for transport of television
- Video teleconferencing
- High-resolution graphics
- Fast fax

Only H0 and H11 are used in the North American version of ISDN (H12 is used in the European version). The H21 and H22 channels were intended for broadband ISDN (BISDN); H4 was designed to transfer high-definition televsion (HDTV). However, it appears that HDTV will use standard digital compression (MPEG2) at approximately 18 Mbit/s rather than 135 Mbit/s (Table 8.4).

Primary rate ISDN may be structured in several ways other than $23B + D$. Table 8.5 lists the synchronous primary rate structures.

Table 8.4
H Channel Data Rates

Component channel	Data rate (kbit/s)	Multiples of B channel	Multiples of H0 channel rate
H0	384	6	1
H11	1536	24	4
H12	1920	30	5
H21	32,768	512	—
H22	44,160	690	115
H4	135,168	2112	352

Table 8.5
ISDN Synchronous Primary Rate Structures

1.544-Mbit/s primary rate channel structures	2.048-Mbit/s primary rate channel structures
$23B + D$	$30B + D$
$3H0 + 5B + D$	$NB + D(1 \leq n \leq 29)$
$3H0 + D$	$5H0 + D$
$3H0 + 6B$	$nH0 + mB + D(0 \leq n \leq 5, 0 \leq m \leq 30, 6n + m < 30)$
$24B$	$31B$
$4H0$	$5H0$
$H11$	$H12 + D$

User Interface for BRI

The (user-provided) **network termination device (NT1)** or **terminal adapter**, which is the interface between the user and the network, does not use either the AMI or the B8ZS line code, but instead uses a coding scheme **2-binary 1-quaternary (2B1Q)**, which is a four-level code. This coding yields a 160-kbit/s line signaling rate, of which 16 kbit/s is used by the service provider for timing and maintenance functions. The other 144 kbit/s provides the 2 × 64 kbit/s for the two B channels and the 16-kbit/s D channel. Figure 8.13 is an example of 2B1Q coding. This signaling only travels between the NT1 and the central-office **exchange terminator/line terminator (ET/LT)** so that dc buildup is not a problem. The four-level signaling, or "quat," uses $\pm 1 - V$ and $\pm 3 - V$ levels. (A different signaling format is used in the ITU-T standard.)

Note the flexibility of BRI. One can use both B channels for voice, both for separate data channels, or one for each. One can also multiplex the two B channels as a single 128-kbit/s data channel. The D channel is, for the most part, available for user data, although control signaling takes priority over user data. It is this priority override feature that requires a protocol such as X.25. Another interesting feature of BRI is that the telephone line connecting the NT1 to the line terminator may be either two-wire or four-wire. Hence, standard telephone wiring may be used.

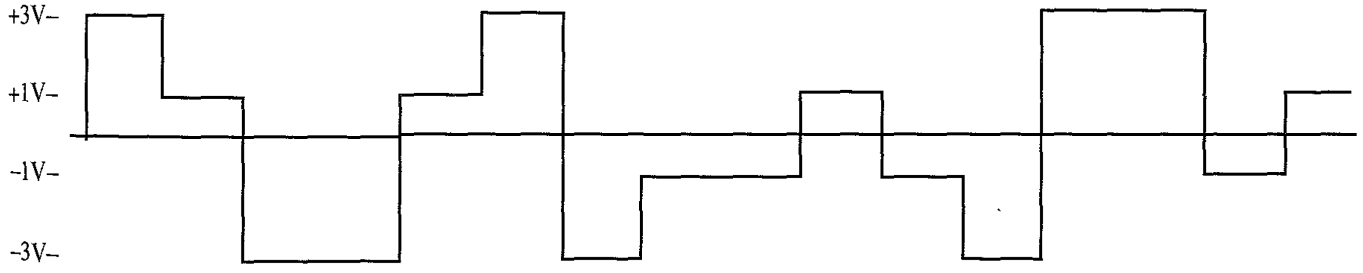
Applications of ISDN

Basic rate ISDN is replacing the modem, albeit (very) slowly. As the telephone network implements SS7 (Signaling System 7) network switches (which are required to transport ISDN signals), users will be able to transmit data at 64-kbit/s DS0 speeds (or 128 kbit/s, if desired) without using a modem.

In the development of ISDN there have been numerous user trials which demonstrated possible applications. Most of these involved the use of voice and data channels on the same ISDN line. One of the more interesting applications of BRI has been in compressed interactive video. For interactive video systems operating with H0 rates, three BRI channels may be multiplexed to form an H0.

Specific trials that represent applications of ISDN include:

- *Westinghouse Electric Corporation*: The Communications Systems Division of Westinghouse began in March 1990 to deploy ISDN PRI lines on Northern Telecom DMS 250



Quat	+3	+1	−3	−3	+1	+3	−3	−1	−1	+1	−1	−3	+3	+3	−1	+1
Bits	10	11	00	00	11	10	00	01	01	11	01	00	10	10	01	11

Figure 8.13 2B1Q line code, North American standard

switches. The ISDN links fulfill several roles at Westinghouse, including fast delivery of engineering drawing revisions, which formerly were transmitted via overnight mail. ISDN also provides cost-effective desktop video conferencing.

- *Andersen Consulting*: In November 1989 Illinois Bell and AT&T Network Systems began a demonstration project providing service to Andersen Consulting which uses two BRI Ameritech ISDN Centrex lines and AT&T's Switched Digital International (SDI) service to link its Tokyo office and its Chicago world headquarters. Illinois Bell, in March 1989, became the first local telephone company in the United States to announce a BRI ISDN tariff. Andersen evaluated the link relative to the cost savings of using ISDN over traditional leased-line video-conferencing methods using a pair of 56-kbit/s lines or a T1 line. In the demonstration, the two B channels of one ISDN line were combined into a single 112-kbit/s video channel; voice traffic was interleaved with the video. The result was a "business quality" picture, which is better than compressed video but not television quality. In addition to the video connection, Andersen will use a second BRI line for Group IV facsimile service. Because AT&T does not have a BRI ISDN service, the two ISDN Centrex lines had been rate-adapted to 56 kbit/s, the speed of AT&T's SDI service. The two were then carried from an AT&T 5ESS digital central-office switch in downtown Chicago over separate 56-kbit/s facilities to Tokyo, where they were carried by Kokusai, Denshin Denwa Co. Ltd. (KDD), the Japanese international carrier. KDD passed the signals to Nippon Telegraph and Telephone Corporation (NTT) and then to Andersen's Tokyo offices, where a terminal adapter will convert the circuits to the standard 64-kbit/s ISDN rate.
- *Harah's Club*: Harah's Reno began testing ISDN applications in April 1989. Its goal was to cut down the delays involved in room registration, thus speeding its customers toward their immediate destination, the gaming tables. The test involved using an ISDN circuit to connect an electronic kiosk at the airport with the hotel reservation desk. Upon arrival at the airport, a traveler intending to stay at Harah's would go to the kiosk and enter a room request using a screen menu. A desk clerk from the resort would appear on a large monitor and speak directly to the customer, who could respond by using a hand set at the kiosk. The terminal also contained a magnetic card reader to handle credit cards and a camera that would capture an image of the person and transfer it as a facsimile image to the hotel. When the customer reached the hotel, a doorman, who recognized the person from the fax image, would greet the customer personally at the door, take the luggage, and hand him or her a room key. While the bags were transferred to the assigned room, the customer would be free to proceed straight to the casino. The process was seldom completed as planned, but not because of equipment problems. The problems were in the human element; the system physically scared people and many customers simply ran away when the clerk came on the screen and began talking to them. Approximately 10% of the customers encountering the kiosk actually used it to completion.

8.2.2 Frame Relay

Frame Relay is basically a high-speed packet switched service that utilizes a fixed-frame structure. It is not a switching technology; rather it is an interface specification which describes how the bits appear to the network. The Frame Relay service provider must address a variety of issues or service elements:

1. Addressing
2. Customer access interface
3. Network Provisioning and maintenance
4. Bandwidth management

Addressing uses a **data link connection identifier (DLCI)** to determine the logical links by which frames are transmitted from the originating end to the receive end of the link. There may be a number of DLCIs for each physical interface (see Fig. 8.14). This allows the end users to consolidate several private lines onto one physical circuit. Frame Relay uses a **permanent virtual circuit (PVC)** for each service subscriber. Essentially, a PVC involves a permanent connection to the network from the subscriber's location. The actual circuit path utilized on any transmission, however, will vary from transmission to transmission and, perhaps, among frames in a single transmission. The PVC only guarantees that some path will be available, not that a specific permanent path will be used. The service provider's network switching determines the actual path on a dynamic basis.

The access interface to the customer involves a **user-to-network interface (UNI)**, which specifies the physical interface between the user and the network. The UNI defines the frame format and the system support, such as PVC management. The access interface also defines the frame transmission rates that the user will access. Typically, these are 56 kbit/s or 1.544 Mbit/s; however, other rates are available from 9600 bit/s to DS1. Access may be via a private line, ISDN, or other means.

Network provisioning and management involves billing for a usage-sensitive service. Frame Relay is not a switched service (a single PVC is a point-to-point circuit). Since it uses virtual circuits which are switched on a dynamic basis, the customer is only charged for actual transmission time. Note, however, that **switched virtual circuits (SVCs)** are possible and will also be available with Frame Relay. The provisioning and management function also provides

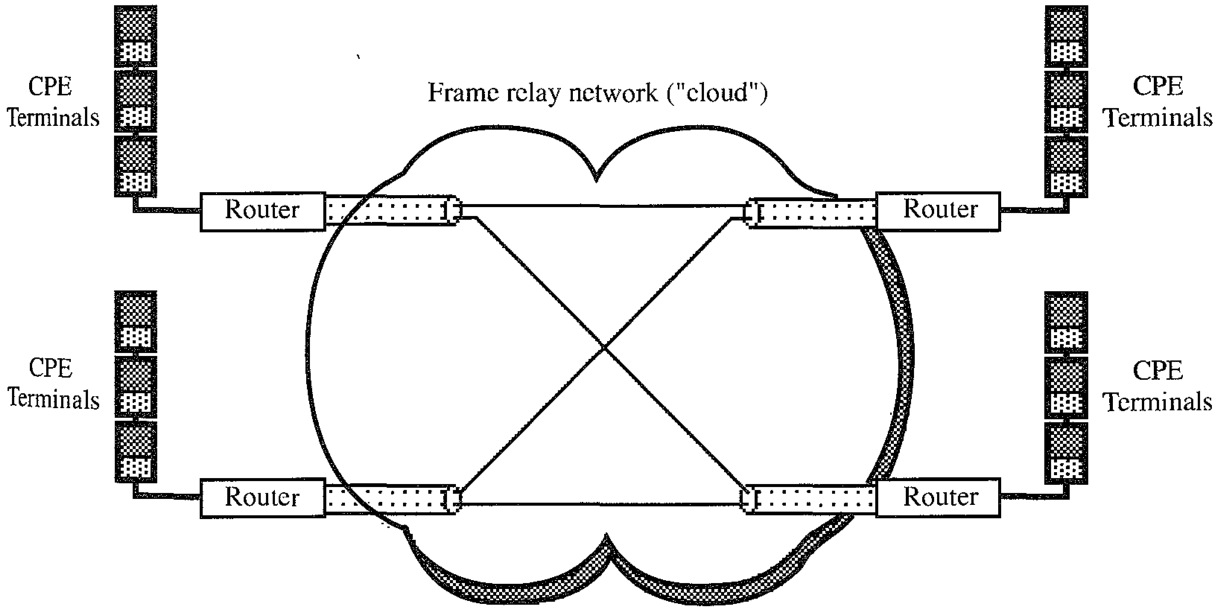


Figure 8.14 Frame Relay conceptual network—multiple logical links via a single physical link

customers with network management options, such as the capability to establish transmission priorities.

Bandwidth management provides service options to the customer. These options include:

1. Basic service, which provides bandwidth on demand.
2. **Committed information rate (CIR)** service, which defines a virtual channel of a given guaranteed bandwidth through the Frame Relay network.
3. Burst rates up to the bandwidth of the physical circuit "pipe."

There are several advantages to Frame Relay and the PVC concept as opposed to standard packet switching or other private-line services. These include:

1. Frame Relay is more cost-effective for bursty traffic. If the bandwidth of a private line is inadequate, application throughput encounters bottlenecks. If private-line bandwidth is excessive, idle facilities waste money.
2. Frame Relay reduces the number of physical line terminations. For N sites with full private-line connectivity one has N line terminations for the first site, $N - 1$ for the second (the second is already connected to the first), $N - 2$ for the third, and so on, or

   $$T_N = \sum_{j=1}^{N} j = \frac{N(N-1)}{2}$$

   required line terminations. Frame Relay requires only N line terminations.
3. Frame Relay provides different access rates for different user sites. This allows lower access rates at network branches and higher access rates at data centers. Moreover, the network provides the necessary adaptation of different rates.
4. Since site-to-site connections are logical rather than physical, changes in the network configuration are easily accomplished. Both UNI and NNI (**network-to-network interface**) standards are in place.
5. Access to the Frame Relay network is relatively simple (Fig. 8.15).
6. Frame Relay provides for "multicasting," which is a group broadcast capability. A frame sent to the group DLCI 27 is replicated for transmittal on all of the group members' DLCI links (Fig. 8.16). A DLCI may belong to several groups.

Frame Structure

The frame structure is relatively simple. It consists of two 1-octet (byte) header/trailer flags, which define the beginning and the end of a specific frame, a 2-octet DLCI or address, an N-octet user message (information field), and a 2-octet cyclic redundancy **frame check sequence (FCS)**, as shown in Fig. 8.17.

The advantages of Frame Relay and this frame structure over standard packet switching are:

1. Improved network transmission facilities
2. More intelligence in the endpoints
3. Endpoints responsible for error recovery
4. Higher data rates available

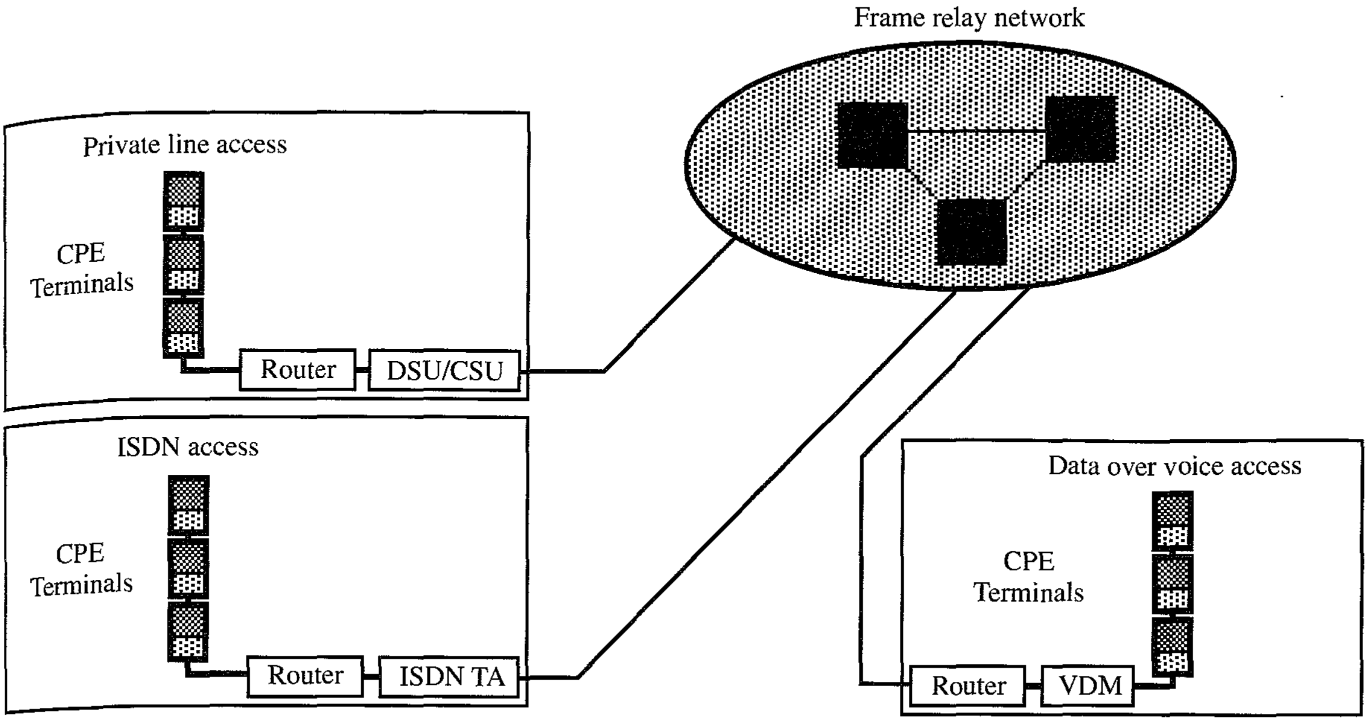

Frame relay network
Private line access
CPE Terminals
Router
DSU/CSU
ISDN access
CPE Terminals
Router
ISDN TA
Data over voice access
CPE Terminals
Router
VDM

Figure 8.15 Customer access service elements. TA—terminal adapter, VDM—voice/data multiplexer.

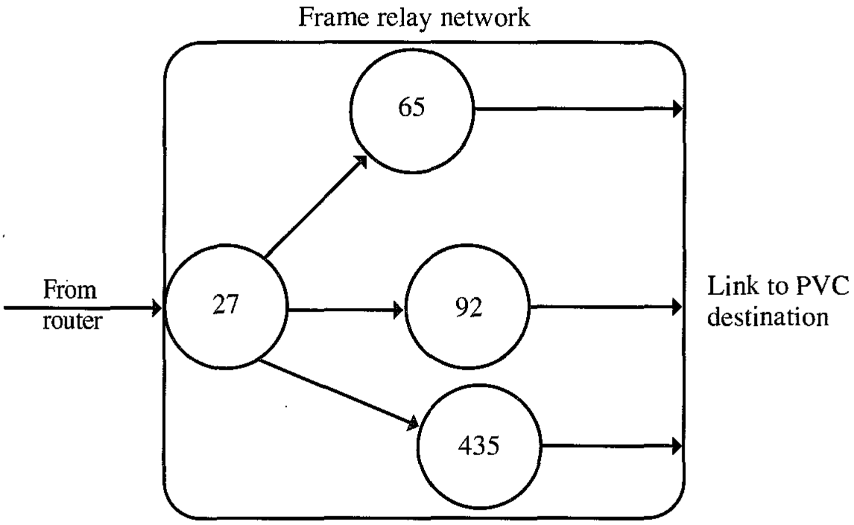

Frame relay network
65
From router
27
92
435
Link to PVC destination

Figure 8.16 Multicasting.

The destination addressing in the DLCI octets is simple. The DLCI indicates which channel the frame is on, hence identifying the PVC. It also contains the **forward explicit congestion identifier (FECI)** and the **backward explicit congestion identifier (BECI)**, which are congestion control flags, and the **discard eligibility (DE)** flag, a frame loss priority flag that identifies if a frame may be discarded in the event of congestion (Fig. 8.18). **C/R** is the **command/response** flag, and **EA** is the **extended address**. The DE and C/R are user defined; the FECN and BECN (the FECI and BECI) are network defined. The DLCI occupies the six most significant bits of the second octet and the four most significant bits of the third octet of a frame. There are 991 valid DLCI permanent virtual circuit addresses. The DLCI defines

Flag (01111110)	DLCI	Information field (variable length)	FCS	Flag (01111110)
1 octet	2 octets	N-octets	2 octets	1 octet

Figure 8.17 Frame structure.

the link between the user and the network, that is, it has local significance only; *only* the user device and the network interface use a particular DLCI . The DLCI must be changed from send-DLCI to receive-DLCI to identify the receive channel to the destination.

Management of Permanent Virtual Circuits

One of the more significant features of Frame Relay is the PVC management. Before a user device transmits data across the UNI, it has the option to query the network to verify the integrity of the PVC link. The UNI can send a series of PVC management requests to identify the available PVCs. Link "keep-alive" signals are continually sent between the user Frame Relay CPE (customer premise equipment) and the network. The PVC management messages, however, only pass status information between the CPE device and the network port to which it is connected.

Another critical part of PVC management is congestion control. It is vital that the network have some means to control the flow of data. The FECN, BECN, and DE provide this control. As the network becomes congested, it sets the FECN and BECN bits so that the attached user devices are aware that congestion is occurring. Theoretically, at least, the user devices can then reduce their flow of data until the congestion bits are no longer set. If congestion continues, frames that are identified by the DE bit as "discardable" are discarded by the network. In reality, however, most routers have no way to pass flow control information back across the local user network to user sending devices. Hence, even though the network can set FECN and BECN bits, such routers do not even check these bits. Note that the CPE sending devices are *not* Frame Relay specific devices. They are standard terminals, personal computers, and the like, and have no concept of Frame Relay protocol specifics. Thus, the router is not able to determine the discard eligibility of the data and, hence, the DE bit setting. Hence, congestion decisions are, in practice, left to the network. Frame Relay **packet assembler/dissemblers (PADs)** will be able to respond appropriately to congestion bits and control the flow of attached devices.

The devices that allow appropriate management of the PVCs and the Frame Relay network are the UNI and the NNI. These interfaces are simple and universal so that service provider networks (e.g., the regional Bell operating companies and other interexchange carriers) can be readily interconnected. By maintaining the Frame Relay protocol across network boundaries, protocol conversion is avoided.

DLCI (msb)						C/R	EA
2^9	2^8	2^7	2^6	2^5	2^4		

DLCI (msb)				FECN	BECN	DE	EA
2^3	2^2	2^1	2^0				

Figure 8.18 DLCI structure.

Frame Relay Services

The "vanilla" flavored Frame Relay service provides only bandwidth on demand. The CIR service provides a virtual channel of specified bandwidth through a Frame Relay network (Fig. 8.19). It provides a minimum throughput with the ability to burst at higher rates. CIR can be considered to be a guaranteed minimum if the system is conservatively designed (e.g., two 32-kbit/s CIRs on a 64-kbit/s access line). It can also be considered to be a statistical guarantee if the service provider allows oversubscription (e.g., four 32-kbit/s CIRs with 50% average utilization on a 64-kbit/s access line). In contrast, X.25 packet-switched networks typically do not guarantee a specified throughput.

In Fig. 8.19 B_c is the maximum amount of data, in bits, that the network guarantees to transfer in time T; B_e is the maximum amount of uncommitted data that the network will attempt to deliver in time T_c and can be set at a rate up to the port rate. Hence, with CIR = 32 kbit/s, the network (via a DS0 64-kbit/s port rate) will guarantee data transfer in 2 seconds. As a practical matter, the net user data rate will probably fall below the CIR due to the bursty nature of the traffic.

8.3 BROADBAND DIGITAL COMMUNICATIONS: SONET

Several broadband communication technologies use a fixed frame/cell structure which is similar to, but somewhat more complex than, the frame structure of T1 or Frame Relay. We shall approach these technologies first from the transport (transmission) and, second, from the switching viewpoint.

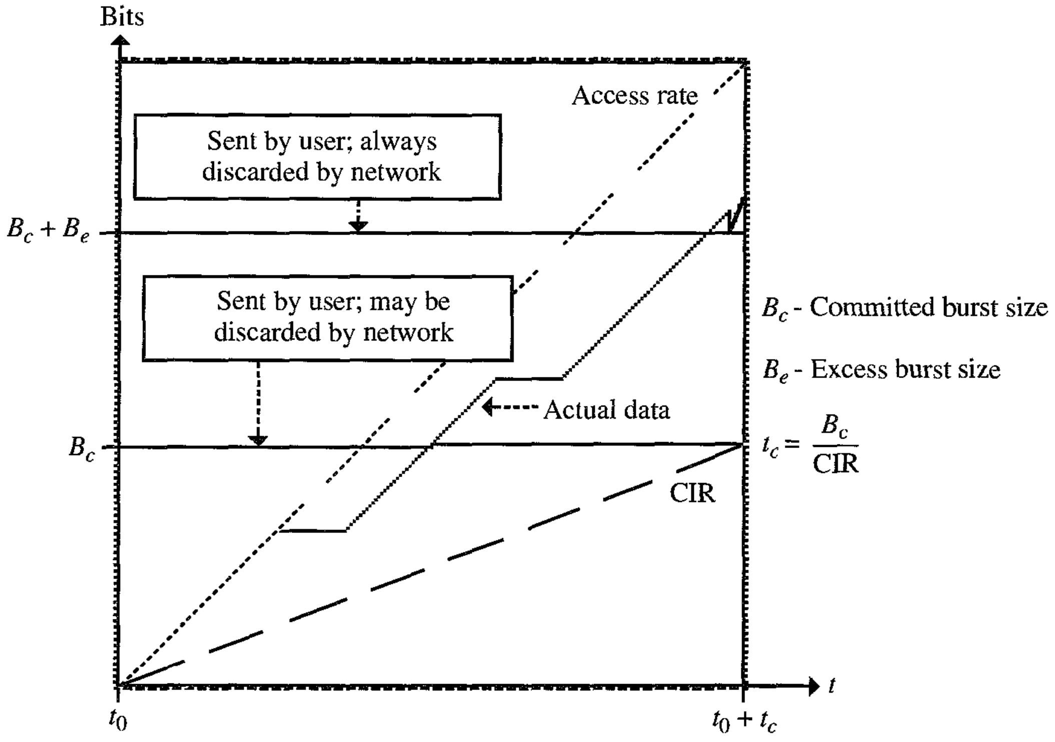

Bits
Access rate
Sent by user; always discarded by network
$B_c + B_e$
Sent by user; may be discarded by network
Actual data
B_c
CIR
B_c - Committed burst size
B_e - Excess burst size
$t_c = \frac{B_c}{\text{CIR}}$
t_0
$t_0 + t_c$
t

Figure 8.19 Committed Information Rate.

One of the fairly recent developments has been the **Synchronous Optical NETwork (SONET)**. SONET is a fiber-optic transmission protocol that provides synchronous transport by means of a very clever and simple, at least conceptually, scheme. SONET is used primarily for network transport of broadband communications (51.84 Mbit/s or greater) between and among carrier (telephone company) switching nodes. The key to its great utility lies in its cell structure and its synchronization protocol.

The standards were developed by BELL COmmunications REsearch (Bellcore), the joint research arm of the regional Bell operating companies, for the following purposes:

- Synchronous fiber-optic networking
- Efficient add/drop multiplexing (ADM)
- Compatibility of equipment among manufacturers of equipment
- Robust fiber rings (survivable fiber-optic rings which can tolerate a single cut in the fiber path)
- Support of new services such as ATM
- Enhanced network management capabilities (called OAM&P for operations, administration, maintenance & provisioning)

These standards also define a sequence of optical transmission rates [optical carrier or (OC)] or levels (Table 8.6a) and their corresponding electrical equivalents [synchronous transport signal (STS)] (Table 8.6).

We should observe that everything is done at the electrical (STS) level prior to the electric-to-optic conversion for transmission. This includes multiplexing, switching, signal grooming, and so on. The higher OC levels (OC-96 and above) are typically accomplished by means of **wave division multiplexing (WDM)** in which multiple optical signals, at different wavelengths, are multiplexed onto a single fiber. This multiplexing technology was made possible by a device called the **distributed feedback (DFB) laser**. The DFB laser has an extremely narrow optical band-spread from its center frequency and, hence, allows several optical signals to be transmitted simultaneously (from individual lasers) at different frequencies on a single fiber without their side-bands interfering with the other frequencies.

Table 8.6
Synchronous Transport Signal Hierarchy

Optical carrier		Synchronous transport signal	
Level	**Rate**	**Level**	**Rate**
OC-1	51.84 Mbit/s	STS-1*	51.84 Mbit/s
OC-3	155.52 Mbit/s	STS-3	155.52 Mbit/s
OC-9	466.56 Mbit/s	STS-9	466.56 Mbit/s
OC-12	622.08 Mbit/s	STS-12	622.08 Mbit/s
OC-18	933.12 Mbit/s	STS-18	933.12 Mbit/s
OC-24	1244.16 Mbit/s	STS-24	1244.16 Mbit/s
OC-36	1866.24 Mbit/s	STS-36	1866.24 Mbit/s
OC-48	2488.32 Mbit/s	STS-48	2488.32 Mbit/s
OC-96	4976.64 Mbit/s	STS-96	4976.64 Mbit/s
OC-192	9953.28 Mbit/s	STS-192	9953.28 Mbit/s
		STS-xxc	Concatenated

*STS-1 bit rate = 810 bytes/frame × 8 bits/byte × 1 frame/125 μs = 51.840 Mbit/s.

8.3.1 SONET Multiplexing

Figure 8.20 illustrates the basic SONET multiplexing scheme. Any type of telecommunications service can be accepted by SONET by means of devices called **service adapters**. Services include voice, high-speed data, digital video, and so on. The service adapter "maps" the specific signal into the payload or information portion of the STS-1 or, in the case of DS1 signals, for example, into a **virtual tributary** (**VT**). All inputs are eventually mapped into the cell structure of the STS-1.

8.3.2 SONET Frame Format

The SONET frame format uses the transmission rate of the STS-1 as its fundamental building block. Higher level (STS-n) signals are simple multiples of the STS-1, that is, there is no additional overhead as in the DS-n hierarchy; STS-n = n × STS-1. The frame is divided into the transport overhead and the **synchronous payload envelope** (**SPE**). The SPE, in turn, is divided into two parts, the STS path overhead and the payload (user-transmitted information). Once the payload is multiplexed into the SPE, it can be transported through the SONET

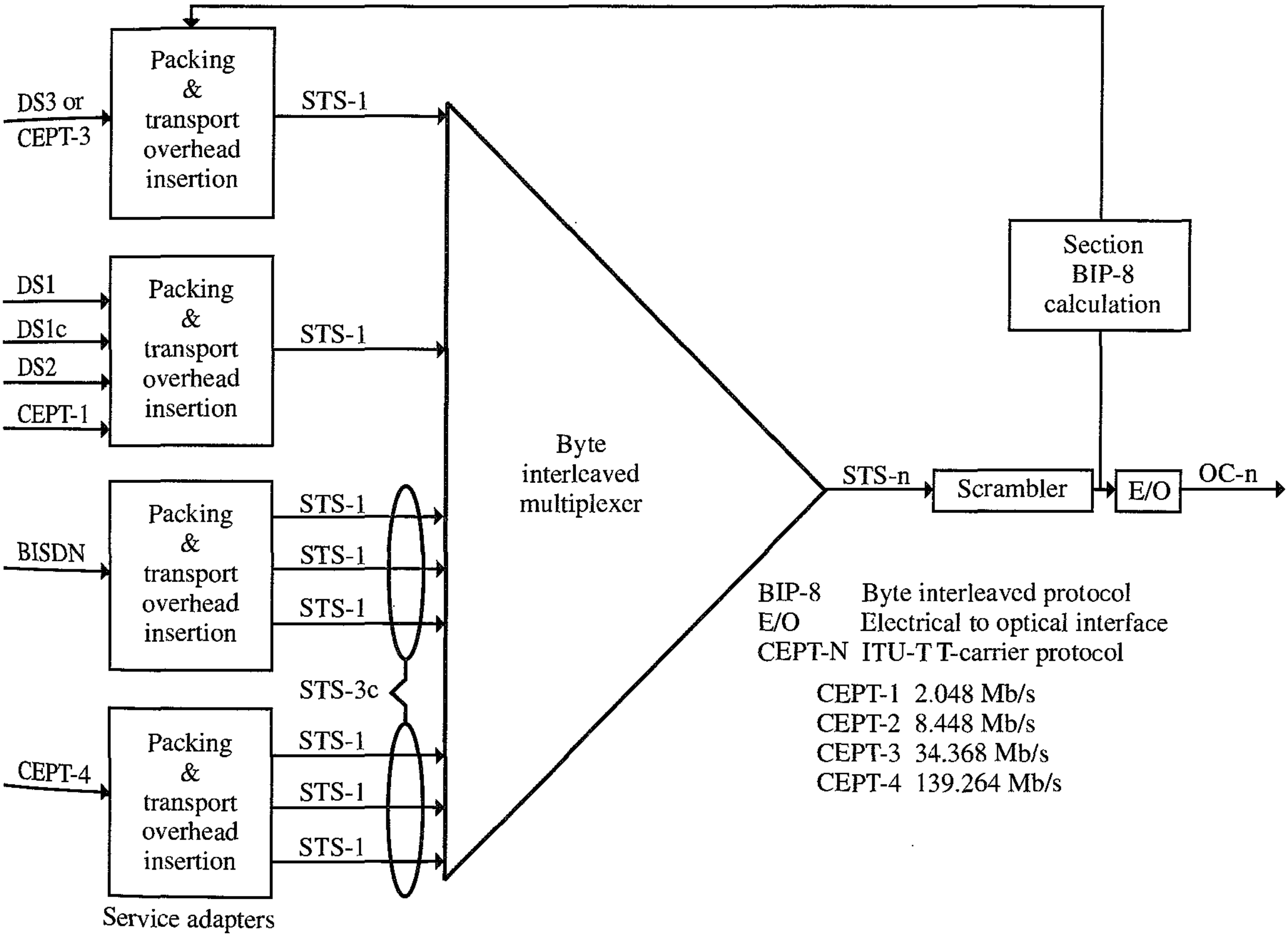

DS3 or
CEPT-3
Packing & transport overhead insertion
STS-1
DS1
DS1c
DS2
CEPT-1
Packing & transport overhead insertion
STS-1
BISDN
Packing & transport overhead insertion
STS-1
STS-1
STS-1
STS-3c
CEPT-4
Packing & transport overhead insertion
STS-1
STS-1
STS-1
Service adapters
Byte interleaved multiplexer
Section BIP-8 calculation
STS-n
Scrambler
E/O
OC-n
BIP-8 Byte interleaved protocol
E/O Electrical to optical interface
CEPT-N ITU-T T-carrier protocol
CEPT-1 2.048 Mb/s
CEPT-2 8.448 Mb/s
CEPT-3 34.368 Mb/s
CEPT-4 139.264 Mb/s

Figure 8.20 SONET multiplexing.

network without having to be demultiplexed at intermediate network nodes, even at add/drop multiplexers, as it would in traditional non-SONET transport. Hence, SONET is said to be a transparent, or service-independent, transmission scheme.

The STS-1 has a payload capacity of 672 voice (DS0) signals, 28 DS1s, or one DS3. Figure 8.21 illustrates the SONET frame structure. Note that the 9 × 90-byte frame is 125 μs in duration, so that the bit rate is

$$9 \times 90 \frac{\text{bytes}}{\text{frame}} \times 8 \frac{\text{bits}}{\text{byte}} \times 8000 \frac{\text{frames}}{\text{second}} = 51.840 \frac{\text{Mbit}}{\text{second}} = \text{STS-1}$$

The order of transmission is left to right, row by row. The most significant bit of each byte is the first bit transmitted. Observe that the 125-μs frame length allows SONET to accommodate multiplexed DS-n frames, all of which transmit in 1/8000 of a second, or 125 μs.

The transport overhead is divided into **section overhead (SOH)** and **line overhead (LOH)**. The SOH is related to transmitter and receiver functions:

- Framing
- Line coding
- Section (TX/RX) error monitoring
- Section maintenance
- Local orderwire (a voice line used in troubleshooting TX/RX problems)
- Performance monitoring of STS-n signal

The LOH is related to:

- Synchronization
- Multiplexing
- Line error monitoring
- Line maintenance

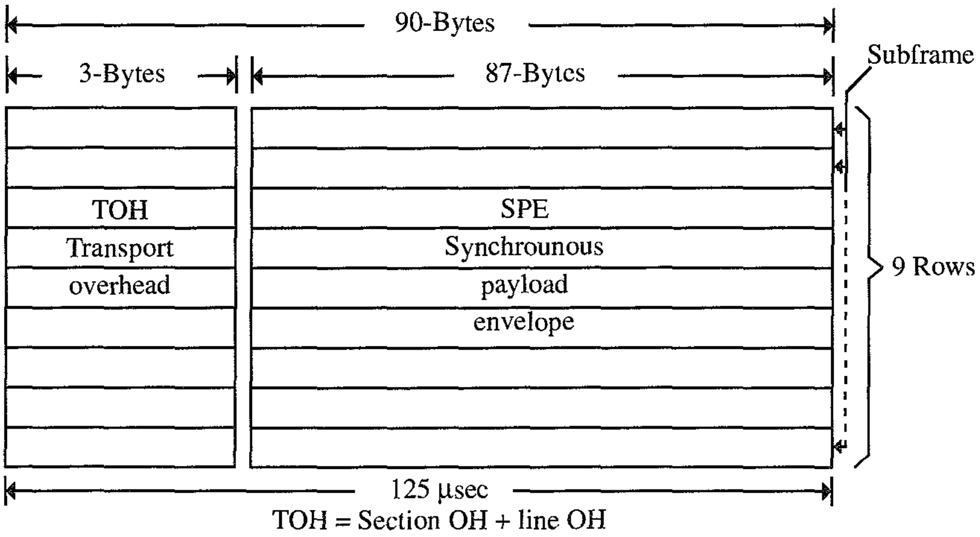


Figure 8.21 SONET frame structure.

- Protection switching
- Line function orderwire (a voice line used in troubleshooting line problems)
- Performance monitoring of individual STS-1s
- Line **far-end receive failure (FERF)** indication
- Pointer to start the synchronous payload envelope

The first byte of each row of the SPE is **path overhead (POH)**, which contains information related to the payload mapping into the SPE, such as:

- Path communications
- Error monitoring
- SPE signal label
- Status and superframe phase indication
- Performance monitoring of the SPE

The POH is relevant to the end-to-end communication of the payload and remains with the SPE until the frame is demultiplexed. The remaining 86 bytes in each subframe contain the payload, for example, a DS3. The DS3 signal has a rate of 44.736 Mbit/s ± 20 parts per million. This yields a maximum of 44.736 Mbits + 895 bits per second and a minimum of 44.736 Mbits − 895 bits per second. Since the SPE contains nine rows of subframes at 8000 frames per second, the maximum data rate supported by the SPE is 44.784 Mbit/s, which accommodates the DS3. If, however, a higher data rate signal is to be carried via SONET, the signal can be concatenated into three or more STS-1s as an STS-3c (155.52 Mbit/s) or a multiple of an STS-3c.

On the other hand, if one wishes to transport a lower rate signal, such as a DS1, CEPT-1, or DS2, via SONET, there are synchronous formats at sub-STS-1 levels. These are the virtual tributaries, of which there are three sizes, as listed in Table 8.7.

Virtual tributaries are multiplexed into an STS-1 as a VT group. A VT-6, three VT-2s, or four VT-1.5s occupy one VT group. A VT group may be of only one type, but seven VT groups may be multiplexed into an STS-1. A significant advantage of SONET is that these VTs may be isolated in the STS-1 data stream and extracted from it without demultiplexing the STS-1 since the virtual tributaries are "neatly stacked" in the STS-1. Note, however, that "bit stuffing" may be required in multiplexing virtual tributaries.

8.3.3 Pointers

Perhaps the most significant feature of SONET is the **pointer**. SONET uses pointers to compensate for frequency and phase variations of the various multiplexed signals; that is, the pointer is the means by which transmissions are synchronized. They allow transport across plesiochronous boundaries, or network interfaces of carriers using separate network clocks. Otherwise, large (125-μs frame sized) "slip buffers" would be required to synchronize transport across these boundaries.

Table 8.7

Type	Transports		VT rate
VT-1.5	1 DS-1	1.544 Mbit/s	1.728 Mbit/s
VT-2	1 CETP-1	2.048 Mbit/s	2.304 Mbit/s
VT-6	1 DS-2	6.312 Mbit/s	6.912 Mbit/s

Pointers provide a dynamic, flexible means to align the payload (user data) in its synchronous payload envelope, independent of the content of the payload. Pointers are dynamic in the sense that the pointer value may be continuously adjusted to track the beginning of the SPE relative to the transport overhead. This allows simple dropping, inserting, or cross-connecting payloads in the network. It also allows simple compensation for signal wander or jitter. If there are any frequency or phase variations between the STS-1 frame and the SPE, the pointer value is automatically adjusted to maintain synchronization.

There are two levels of pointers in SONET. The first level consists of two bytes in the line overhead which identify the beginning of the SPE (which may begin at any point in the payload portion of an STS-1 frame, as shown in Fig. 8.22). The second level is the VT level pointer, consisting of two bytes which identify the beginning of the payload in a VT group.

8.3.4 Applications

Since SONET is modular and payload independent, it provides the network a great deal of flexibility in providing telecommunication services. It supports high-speed packet services, DS1, DS2, and DS3 transport, LAN interconnection/transport, HDTV transport and numerous other services. One of the growing uses of SONET is for the transport of the asynchronous transfer mode (ATM), which we shall discuss in some detail in the next section.

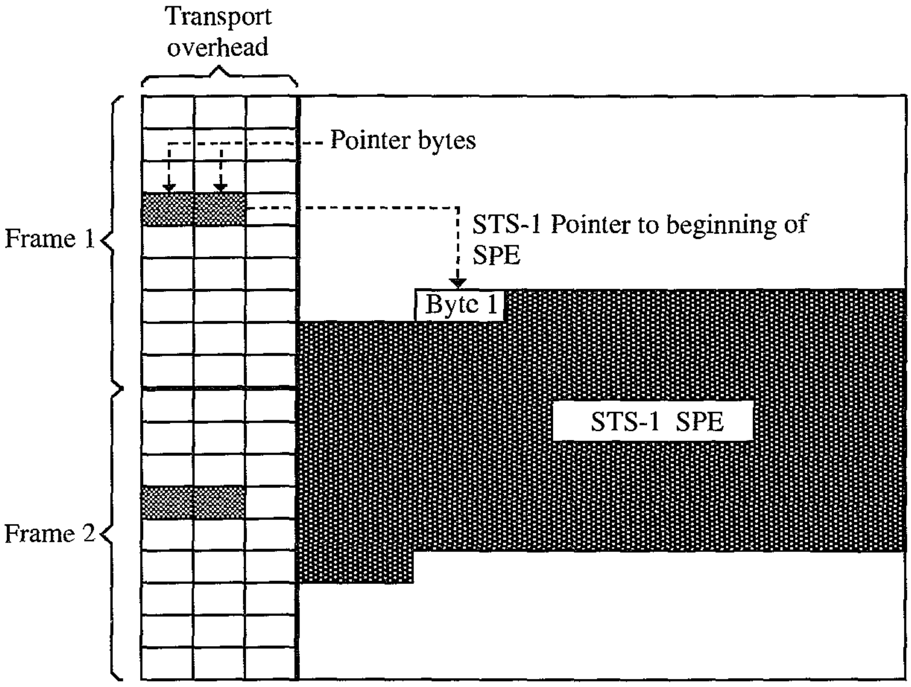


Figure 8.22 Pointer.

8.4 DIGITAL SWITCHING TECHNOLOGIES

We have considered the transport or transmission of digital signals across a network but have not mentioned how signals are switched so as to arrive at their appropriate destinations. A detailed study of digital switching is beyond the scope of this text. There are two modern switching technologies that are fundamental to modern digital communications which deserve mention. One of these is **Signaling System 7 (SS7)**; the other is **asynchronous transfer mode (ATM)**.

8.4.1 Signaling System 7 (SS7)

SS7 is designed for interoffice network switching, that is, switching signals between and among telephone central offices. It utilizes a feature called **common channel interoffice signaling (CCIS)**, which simply means that all control signaling is done on a common channel rather than on a portion of a voice channel. It is a fundamental part of ISDN transport; without SS7 network switches, ISDN cannot be a viable service across the telephone network. Probably this is the reason why the growth of ISDN has been relatively slow. As SS7 switching is integrated into the public switched telephone network, however, ISDN services will continue to grow.

8.4.2 Asynchronous Transfer Mode (ATM)

ATM is a technology that has developed since 1990. Basically, ATM is a broadband multiplexing scheme which allows the multiplexing of widely differing types of digital signals into a common digital stream. It provides the primary mechanism for switching **broadband ISDN (BISDN)**, and, hence, discussion of ATM inherently involves BISDN. The multiplexing is not standard (synchronous) TDM, as is the case in most digital switching schemes. Rather, it is **statistical time-division multiplexing (STATDM)**, which takes advantage of the bursty nature of most communications to provide an overall higher data throughput. Statistical multiplexing, sometimes called **asynchronous time-division multiplexing**, is closely related to packet switching. (Indeed, ATM is sometimes called high-speed packet switching.) In TDM a time slot is assigned to each data source (e.g., a voice channel). In STATDM the time slot is assigned only when the source is active. A STATDM stores and forwards frames of data from node to node. One might think that, at times of heavy traffic, this approach would lead to serious problems of contention for bandwidth. The ATM approach to this contention problem is one of the significant features of the technology.

Unlike traditional packet switching, ATM uses a fixed-frame structure called a **cell**. The ATM cell is 53 **octets** (in ATM bytes are called octets) long, of which 48 octets comprise the message and 5 octets the header. The cell structure was chosen for two reasons. The relatively small cell size reduces queuing delay for high-priority cells in the event of congestion. Fixed-size cells may be switched more efficiently since the switch need not look for an end-of-cell indication. Figure 8.23 describes the basic cell structure.

Figure 8.24 outlines the content of the header at the user interface to the network. The header consists of several portions:

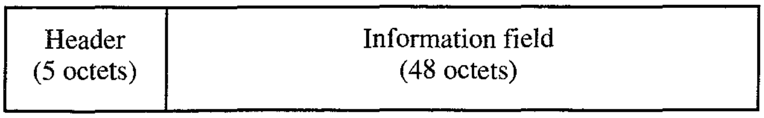

Header
(5 octets)
Information field
(48 octets)

Figure 8.23 ATM cell.

- **Generic flow control (GFC)**: This is a flag, which allows the bidirectional control of cell flow (data) to the network in the event the network becomes congested. It allows the user to define the traffic characteristics of the data that the network is to accommodate.
- **Virtual path identifier (VPI)**: This defines the virtual network path that will be used for routing cells across the network.
- **Virtual channel identifier (VCI)**: This identifies the specific virtual channel/circuit (set of time slots) on which the cell will be carried between the point where the VCI is assigned and the point at which it is terminated or translated to a new VCI.
- **Payload type (PT)**: This 1-bit field specifies whether the information carried in the information field is user data or network data.
- **Cell loss priority (CLP)**: This 1-bit field provides information to the network as to whether or not the cell may be discarded in the event of network congestion—CLP = 1 indicates that the cell may be discarded; CLP = 0 indicates that the cell may not be discarded unless no alternative exists.
- **Header error control (HEC)**: A CRC field using the check polynomial

$$G(X) = X^8 + X^2 + X + 1$$

It is used for error detection and provides capability for single-error correction and very low probability of undetected burst errors of magnitude greater than 1.

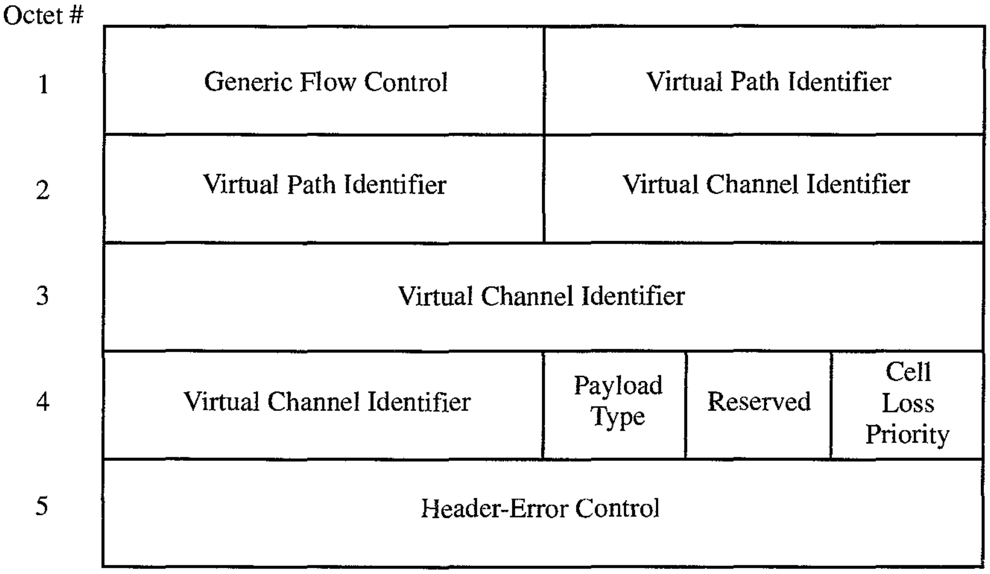

Octet #
1
2
3
4
5
Generic Flow Control
Virtual Path Identifier
Virtual Path Identifier
Virtual Channel Identifier
Virtual Channel Identifier
Virtual Channel Identifier
Payload Type
Reserved
Cell Loss Priority
Header-Error Control

Figure 8.24 ATM cell header field (user-to-network interface).

Figure 8.25 outlines the content of the header at the network-node interface. Let us consider the concept of virtual path and virtual channel somewhat further. Virtual means that the connection is via a time slot as opposed to a fixed pair of wires. A **virtual channel connection (VCC)** connects two endpoints by means of the concatenation of a series of **virtual channel (VC)** links. A virtual channel link is identified by a VCI. A new VCI is assigned whenever a virtual channel link is switched. A **virtual path connection (VPC)** consists of the concatenation of a series of **virtual path (VP)** links, each consisting of a group of virtual channel links. Each virtual path link is identified by a VPI, which is common to all virtual channels in the virtual path link. Within the VPC, each VCC is identified by a unique VCI. VCCs may have the same VCI within different VPCs, however. Figures 8.26 and 8.27 may shed some light on these concepts. Note that when a virtual path is switched, the virtual channels remain unchanged. However, whenever a virtual channel is switched, the virtual path is also switched.

A somewhat "fishy" analogy of the VC/VP concept is as follows. Suppose a seafood wholesaler (the user) in Seattle wishes to ship fresh salmon to a Chicago restaurant (the terminal). He packages the fish in crates (cells) and ships them by air. The crates are assembled on a palette (virtual channel). On the first shipment, the plane goes first to Minneapolis (virtual path 1) and then to Chicago (virtual path 2), the stopover in Minneapolis representing a virtual path switch. The next shipment goes first to Salt Lake City (virtual path 1) and then to Denver (virtual path 2), where the crate (virtual channel 1) is transferred (virtual channel switched) to a palette (virtual channel 2) on an airplane bound for Chicago (virtual path 3). The virtual paths used in both cases terminate in Chicago, but the routing was quite different.

In our analogy, the salmon might represent any of a large number of communication services: BISDN, compressed video, voice (in the form of multiplexed DS1s), high-speed data, high-resolution images, and so on, anything that can be transmitted in an ATM cell. The cell loss priority structure and the flow control features of ATM allow it to transport multiple services without concern over cell loss in critical applications, but with economy of cell transport due to the statistical nature of the multiplexing.

<table>
<tr><th>Octet #</th><th colspan="4"></th></tr>
<tr><td>1</td><td colspan="4">Virtual Path Identifier</td></tr>
<tr><td>2</td><td>Virtual Path Identifier</td><td colspan="3">Virtual Channel Identifier</td></tr>
<tr><td>3</td><td colspan="4">Virtual Channel Identifier</td></tr>
<tr><td>4</td><td>Virtual Channel Identifier</td><td>Payload Type</td><td>Reserved</td><td>Cell Loss Priority</td></tr>
<tr><td>5</td><td colspan="4">Header-Error Control</td></tr>
</table>

Figure 8.25 ATM cell header field (network-to-node interface).

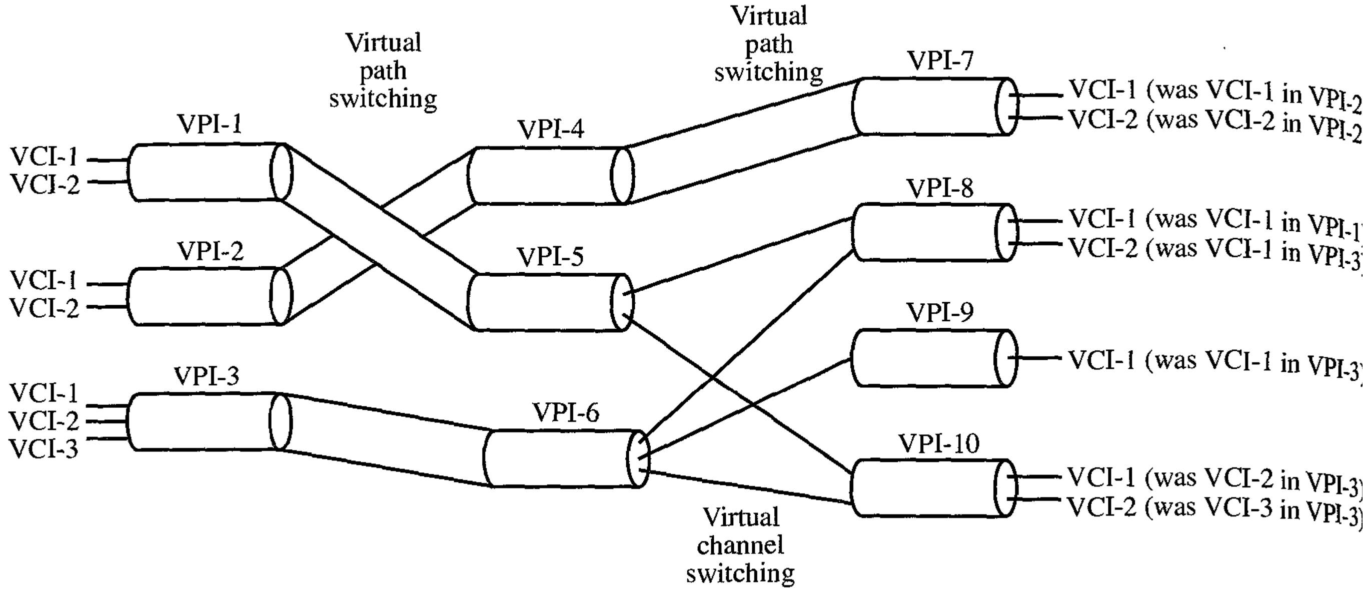

Virtual
path
switching
Virtual
path
switching
VPI-1
VPI-2
VPI-3
VPI-4
VPI-5
VPI-6
VPI-7
VPI-8
VPI-9
VPI-10
VCI-1
VCI-2
VCI-1
VCI-2
VCI-1
VCI-2
VCI-3
VCI-1 (was VCI-1 in VPI-2
VCI-2 (was VCI-2 in VPI-2
VCI-1 (was VCI-1 in VPI-1
VCI-2 (was VCI-1 in VPI-3
VCI-1 (was VCI-1 in VPI-3
VCI-1 (was VCI-2 in VPI-3
VCI-2 (was VCI-3 in VPI-3
Virtual
channel
switching

Figure 8.26 ATM virtual connection relationships.

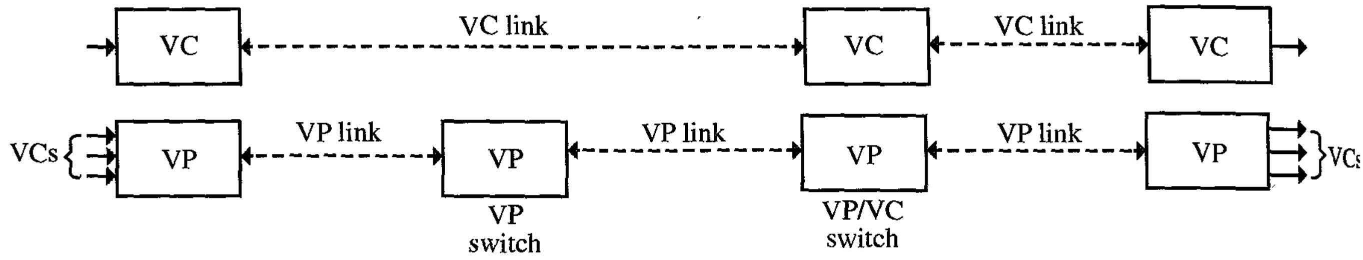

VC
VC link
VC
VC link
VC
VCs
VP
VP link
VP
VP link
VP
VP link
VP
VCs
VP
switch
VP/VC
switch

Figure 8.27 VP/VC link structure.

ATM Structure

ATM is structured in layers, as are many data transmission protocols. The ATM protocol model is pictured Fig. 8.28.

The **physical layer** is concerned with the physical transmission medium and is divided into two sublayers:

1. The **physical medium sublayer** provides the actual transmission and reception of signals over the physical transport medium (e.g., SONET).
    - Line coding.
    - Bit timing.
    - Generation and detection of the transmitted signal.
2. The **transmission convergence sublayer** provides the means for converting an arbitrary bit stream entering the switch into ATM cells.
    - Generation and recovery of transmission (e.g., SONET) frames to carry ATM cells in frame based ATM and mapping the ATM cells to and from the transmission frames.

- Insertion of synchronization markers in stream-based ATM .
- Identification of cell endpoints.

The **ATM layer** defines the protocols for:

- Cell multiplexing to provide the capability to multiplex cells from multiple virtual paths or virtual channels and demultiplexing to direct individual cells to specific virtual channels of virtual paths.
- Translation of VCIs/VPIs to new values at ATM switching or cross-connect nodes.
- Generation of cell headers for attachment to the information received from higher layers and removal of the cell header before transmitting the information to a higher layer.
- Management of access and information flow at the UNI (generic flow control).

The **ATM adaptation layer** (**AAL**):

- Resolves disparities between the service provided by the ATM layer and the service requested by the user. It *adapts* the user data to the ATM data stream.
- Converts user information data streams to ATM cell format.
- Handles transmission errors, lost cells, misinserted cells, and cells with errors.
- Controls user data flow in accordance with user-specified quality of service requirements.
- Contains a **convergence sublayer**, which provides functions that may be required to support specific AAL applications.
- Contains a **segmentation/reassembly sublayer**, which packs the information received from the convergence sublayer into cells for transmission and unpacks received cells.

Up to this point we have only mentioned that ATM can support many varied services. The AAL provides four service classifications for the transport of constant- or variable-bit-rate data streams in connection or connectionless modes and with the option of maintaining timing synchronization between sender and receiver. These service classifications are labeled Class A through Class D and are outlined in Fig. 8.29.

ATM Service Classes

Class A Service: Class A service supports a constant-bit-rate, connection-oriented service with timing synchronization between sender and receiver. Services that require Class A include DS1 and DS3 circuits, BISDN, and constant-bit-rate audio or video. Switched digital video services such as video-on-demand (VOD) or near-video-on-demand (NVOD) will require

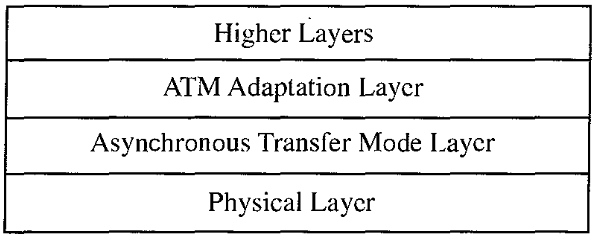


Figure 8.28 ATM protocol model.

	Class A	Class B	Class C	Class D
Bit rate	Constant	Variable		
Connection mode	Connection oriented			Connectionless
Timing syncronization sender/receiver	Required		Not required	

Figure 8.29 Service classifications provided by ATM adaptation layer.

Class A service with constant-bit-rate, compressed, digital video streams. It should be noted that ATM was not designed for voice services, although voice may be carried in DS1 or DS3 multiplexed data streams in Class A service.

Class B Service: Class B service supports variable-bit-rate, connection-oriented service with timing synchronization between sender and receiver. A typical service requiring Class B might be variable-bit-rate digital video for video conferencing.

Class C Service: Class C service supports variable-bit-rate, connection-oriented service with no timing synchronization required between sender and receiver. A typical example of a service that would require Class C would be a connection-oriented data transfer which internally defines the start and stop of the data stream.

Class D Service: Class D service supports variable-bit-rate, connectionless service with no timing synchronization required between sender and receiver. An example might be connectionless data transfer between two remote LANs. Another example would be **switched multimegabit data service** (**SMDS**).

ATM Switching

ATM transmits cells over physical links (e.g., SONET) at rates defined by the SONET hierarchy at STS-3 (155.52-Mbit/s) and/or STS-12 (620.08-Mbit/s) rates. (ATM cross-connects exist, which operate at DS1 rates and format the signal in ATM cells.) An ATM switch must accommodate the statistical variations in the cell stream it accepts. It must also be able to operate in a point-to-point, point-to-multipoint, or broadcast mode to support the wide variety of services outlined.

Due to the statistically multiplexed nature of ATM services, an ATM switch must be able to resolve the various conflicting demands for resources among multiple cells that may arrive in the same time slot. More than one cell destined for the same output port may arrive in the same time interval. Moreover, cells destined for different output ports require the same internal switching resources.

The ATM switch must perform its various functions while maintaining an acceptable **quality of service** (**QOS**) to the customer. This QOS includes:

- A very small probability of cell blocking (so that the cell is not transmitted)
- Throughput capabilities to support STS-3 or STS-12 (i.e., hundreds of megabits per second) at a negligible (10^{-11} or better) bit error rate

- A cell loss probability of 10^{-8} to 10^{-11}, depending on the type and class of service
- A cell misrouting rate of 10^{-11} to 10^{-14}, depending on the type and class of service
- A switching delay less than 1 ms
- Variation in switching delay not to exceed 0.1 ms

The ATM performance parameters are outlined in Table 8.8.

ATM Service Categories

In addition to the service classes, ATM also provides four service categories:

1. *Continuous-bit-rate service with reserved* (guaranteed) *peak bandwidth.* This service category also implies no cell loss and low cell delay variation. Associated with Class A service.
2. *Variable-bit-rate service with reserved bandwidth.* The reserved bandwidth in this service category is a function of the peak bit rate, average bit rate, and variance of the bit rate. It provides a guaranteed maximum cell loss rate and cell delay (larger than those of continuous-bit-rate service).
3. *Variable-bit-rate service without reserved bandwidth.* This service category is used when traffic levels are difficult to predict and users can accept occasional network congestion and the associated cell losses.
4. *Reserved burst bandwidth.* The customer selects from a specified list of available peak rates. The user must request permission to send a burst and wait for network "go-ahead" confirmation.

Applications of ATM

Perhaps the most significant application of ATM is in broadband ISDN (BISDN). The two technologies have developed in close parallel. Other applications include video-on-demand, SMDS, video conferencing, and DS3 switching.

ATM Standards

The ATM Forum provides current information on specifications and standards on its web site, www.atmforum.com. It also provides interesting tutorials on ATM.

Table 8.8
ATM Performance Parameters.

Parameter	Definition
Cell loss ratio	Ratio of lost cells to transmitted cells
Cell misinsertion rate	Number of misinserted (misrouted) cells per connection second
Cell error ratio	Ratio of errored cells to number of delivered cells
Severely errored cell block ratio	Ratio of number of severely errored cell blocks to total number of cell blocks
Cell transfer delay	Δt (switching delay of a single observed cell)
Mean cell transfer delay	Arithmetic average of a specified number of cell transfer delays $-\overline{\Delta t}$
Cell delay variation	$\Delta t - \overline{\Delta t}$

8.4.3 Broadband ISDN (BISDN)

Broadband integrated services digital network (BISDN) is an expanded version of ISDN, which offers information transport with a broad range of data rates and traffic parameters. It is intended to integrate a wide range of data communication services over a single network. Services for which BISDN is proposed as a transport medium include:

- HDTV
- Video teleconferencing
- High-speed, high-definition facsimile
- High-speed (multimegabit) data
- Voice services
- LAN-to-LAN communication
- Telemetry
- High-definition images (e.g., medical images at DS3 rates or higher)

Implementation of BISDN involves high-capacity fiber-optic transport and flexible (e.g., ATM) switching and multiplexing based on over 60 standards, technical advisories, and/or recommendations of the American National Standards Institute (ANSI), the European Telecommunication Standards Institute (ETSI), Bell Communications Research (Bellcore, the research arm of the regional Bell operating companies), and the ITU. It is designed to provide a broad variety of digital communication services to both business and residential customers utilizing a very flexible allocation of bandwidth of up to 155.52-Mbit/s STS-3 rates.

The flexibility of BISDN was designed into the concept from the beginning. BISDN connections may be permanent or switched, point to point (between two sites, e.g., LAN-to-LAN connection), point to multipoint (from one site to several sites, e.g., corporate video conference), or broadcast (from one point to everyone else on the network, e.g. television network broadcast). The switched capability offers one of the major keys to BISDN flexibility. It allows on-demand service access, although permanent or reserved service access is also available. In addition, information transfer may be circuit or packet switched, connection oriented or connectionless, duplex (bidirectional), symmetric (same bandwidth in both directions) or asymmetric (higher bandwidth in one direction), or simplex (one direction) via synchronous or asynchronous transfer modes (Fig. 8.30).

Synchronous Transfer

The bandwidth available to the user (at the UNI) is provided in logical channels with fixed bit rates which are multiples of a DS0, another factor contributing significantly to the flexibility of BISDN. The channels are H channels, which were covered in Section 8.2.1. Specifically:

- H21—512 × 64 kbit/s = 32.768 Mbit/s
- H22—690 × 64 kbit/s = 44.160 Mbit/s
- H4—2112 × 64 kbit/s = 135.168 Mbit/s

These channels are synchronously time-division multiplexed into appropriate channel structures and transmitted across the UNI, which is called **synchronous transfer mode (STM)**. Some of the defined channel structures used in STM are:

- 2016B
- 8064B
- H4+4H12+2B+D (16 kbit/s)
- 4H4+16H12+30B+D (64 kbit/s)
- 4H4+2B+D (16 or 64 kbit/s)
- 3H21+nH12+mB+D (16 or 64 kbit/s)

Asynchronous Transfer

The asynchronous channel structures are based on the transfer of an ATM cell, which consists of a fixed number of time slots, each of which accommodates 1 octet per byte of information. Each ATM cell, as we have observed, is composed of a 5-octet header and a 48-octet information field. An asynchronous channel structure is defined as the *synchronous* transmission of a continuous stream of cells at a specified bit rate across the UNI. There are two types of channel structures in ATM:

- *Frame-based*: The cells are packaged into frames, which contain a fixed number of cells, the frame synchronization, and overhead information.
- *Unframed*: The cells are transmitted individually with no frame structure. Synchronization cells may be inserted periodically in the cell stream as required.

ATM may accommodate BISDN rates up to STS-3c.

8.4.4 Switched Multimegabit Data Service (SMDS)

Switched multimegabit data service (SMDS) is defined by Bellcore as a service for the switching and transport of connectionless (packet-type) data. The data, in the form of individual

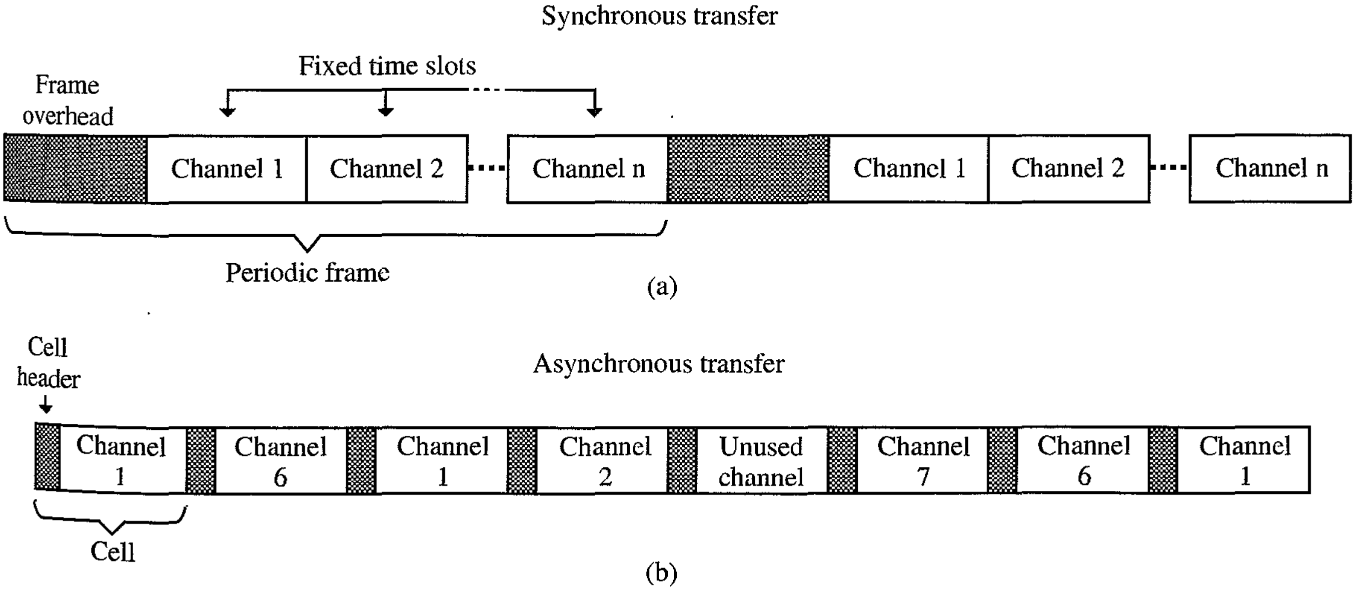

Synchronous transfer
Fixed time slots
Frame overhead
Channel 1
Channel 2
Channel n
Channel 1
Channel 2
Channel n
Periodic frame
(a)
Asynchronous transfer
Cell header
Channel 1
Channel 6
Channel 1
Channel 2
Unused channel
Channel 7
Channel 6
Channel 1
Cell
(b)

Figure 8.30 BISDN transfer modes.

packets addressed to a specific termination point, is switched by the service provider's (e.g., the telephone company's) SMDS switch. In recent years, however, with the advent of ATM/BISDN, SMDS has become a subset of the BISDN (ATM switched) service in the implementation by some service providers.

8.5 BROADBAND SERVICES FOR ENTERTAINMENT AND HOME OFFICE APPLICATIONS

It is interesting to note the impact of entertainment, particularly video-based entertainment, on broadband telecommunications. Entertainment appears to be a major driving factor in the broadband switching and network design of several regional Bell operating companies, for example. One of the reasons for this is the development of digital video compression technology which provides cable-quality video in a 4- to 6-Mbit/s bandwidth. An ATM switched STS-3c, for example, could carry approximately twenty 6-Mbit/s digital video channels or thirty 4-Mbit/s channels for a cable television service provider. A SONET transported OC-48 could carry approximately 336 6-Mbit/s channels, each carried as a VT-6, and 480 channels at 4 Mbit/s using VT-2 groups (with two VT-2s per channel). As a result of the rapidly expanding network transport capabilities and the growing demand for entertainment services, a number of technologies and services are developing.

Technologies such as asymmetric digital subscriber line (ADSL) are also being envisioned as a means for providing broadband channels to homes and offices in order to improve the operation of offices in homes and to enable fast and efficient home shopping.

8.5.1 Video-on-Demand (VOD)

One of the more ambitious services is **video-on-demand (VOD)**. This service is designed to provide the customer access to a large number of first-run movies. The customer, by means of a **set top box (STB)**, can access virtually any movie. The STB decodes the digital signal into analog form and provides menu and control functions to access lists of movie titles, request a specific title, and control the viewing (start, stop, pause, etc.). One can immediately observe that such a service could be quite expensive. For the viewer to have such control over an individual movie would require complex and expensive servers, probably with multiple (perhaps several thousand) copies of popular films. For example, a 90-min film digitally compressed at 4 Mbit/s would require 90 min $\times$ 60 s/min $\times 4 \times 10^6$ bit/s $= 2.16 \times 10^{10}$ bits = 2.7 gigabytes of media storage. One thousand titles, with an average of 100 copies per title, would require approximately $2.7 \times 10^9 \times 10^3 \times 10^2 = 2.7 \times 10^{14}$ = 270 terabytes of storage!

Because of the tremendous storage requirements and costly servers, perhaps at multiple locations, a less costly alternative called **near video-on-demand (NVOD)** has been proposed. NVOD starts an individual film at regular intervals, perhaps 10 min. Instead of several thousand copies for a first-run, highly popular film, only six would be required at each server location. If the viewer needs to leave the television set for a few minutes, he or she would need to wait 10 min at the most to return to the point at which the film was stopped. The STB would automatically determine where to rejoin the film, that is, which of the six transmissions to access.

Note that VOD and NVOD requires two-way communication with the service provider (e.g., cable company or telephone company). The video transmission is referred to as **downstream**, the user control communication is called **upstream**. Note that the bandwidth requirements are substantially smaller in the upstream direction than in the downstream direction. For this reason, such services are called **asymmetric**.

8.5.2 Hybrid Fiber Coax (HFC)

One of the technologies that is transporting VOD/NVOD services, as well as standard and expanded cable services, is **hybrid fiber coax (HFC)**. In such a system, the service provider typically uses fiber-optic transport, often as a fiber loop to provide redundancy, from the cable head end (satellite downlink, SONET terminal, etc.) to a "pedestal" at which the digital signal on the fiber is translated to an analog signal for retransmission to the customer over standard coaxial cable. A pedestal typically serves 500 homes. The roughly 745-MHz bandwidth of the cable is distributed somewhat as follows:

1. 5–50 MHz (35-MHz bandwidth)—upstream (reverse direction).
1. 54–550 MHz (496-MHz bandwidth)—standard cable 6-MHz RF channels (approximately 83).
2. 550–750 MHz (200-MHz bandwidth)—digitized (4- or 6-Mbit/s) video modulating RF signals using 64 QAM (or, eventually, 256 QAM) modulation and **orthogonal frequency division multiplexing (OFDM)**. 64 QAM provides approximately 8-Gbit/s digital bandwidth for 2000 4-Mbit/s channels.

The upstream bandwidth is used for VOD/NVOD control functions, status monitoring (e.g., status of viewing pay-per-view channels), two-way telephony, high-speed data transfer via cable modems, and so on. The standard cable bandwidth is used for basic and premium cable and pay-per-view cable services. The digital bandwidth is used for VOD/NVOD delivery, interactive television services, expanded cable services, multimedia services, and so on. Note that the digital channels require an STB converter to convert the QAM signal into an analog signal. The STB is also required for the upstream control functions and status monitoring.

Observe that HFC networks must be extremely reliable, particularly because of the telephony services they provide. Networks which provide telephone service, for example, are considered to be subscriber "lifeline" (911, etc.) services.

A competing, but similar, technology is called **fiber-to-the-curb (FTTC)**. In this system, the optical fiber is extended to a pedestal, as in HFC. The pedestal, however, serves pockets of 18 to 24 homes with twisted-pair cable at a substantially reduced bandwidth. FTTC is substantially more expensive than HFC.

8.5.3 Asymmetric Digital Subscriber Line (ADSL)

Another technology designed to compete with HFC is **asymmetric digital subscriber line (ADSL)**. It is one of three primary digital subscriber line (X-DSL) technologies:

1. *HDSL* (*high-bit-rate digital subscriber line*): HDSL is a mature DSL technology, which uses two twisted pairs of standard subscriber copper telephone lines. It supports 1.544 Mbit/s up to a distance of 12 kilofeet (kft) and allows voice, data, and compressed video traffic on the circuit. It uses the 2B1Q line code of ISDN.
2. *ADSL*: An emerging technology which allows (roughly) 6 Mbit/s downstream (to the subscriber) and 640 kbit/s upstream for standard telephony, data, and so on. As with HDSL, this is done on standard copper pairs. The distance normally associated with ADSL is 18 kft. In contrast with HDSL, however, ADSL uses a modulated analog carrier.
3. *VDSL* (*very-high-data-rate ADSL*): VDSL is similar to ADSL, but supports (roughly) 26 Mbit/s to 3 kft and 51 Mbit/s to 1.2 kft.

In ADSL the service is carried to a neighborhood pedestal via fiber, as in HFC. The digital signal is not converted to analog at the pedestal, however, but is transmitted over standard copper telephone pairs to a distance of, perhaps, 12 kft (depending on the condition of the copper pairs), typically using **discrete multitone (DMT)** signaling,[5] the ADSL signaling standard. The ADSL system is based on the fact that a common subscriber pair can carry a signal of approximately 1.1 MHz for a limited distance. The DMT line code uses 256 frequency-multiplexed tones with 4.3125-kHz spacing. The tones are rate adaptive, that is, each tone can be individually modulated using 16 QAM (or, as necessary, a lower rate modulation scheme) up to a rate of 32 kbit/s for a theoretical total digital bandwidth of 8.192 Mbit/s. This theoretical limit can rarely be approached in practice due to, what one might expect, common problems associated with copper cable bundles typically used in the telephone subscriber line distribution system. Typical transmission impairment problems include (Fig. 8.31):

- Normal line attenuation.
- Ingress of RF interference. This includes AM stations and 160-m amateur radio transmissions. The power and duration of interfering stations depend on the characteristics of the source and its proximity to the ADSL line.
- Bridged taps. These are unterminated stubs of twisted-pair cable which result in insertion loss and significant "notches" in the 1.1-MHz spectrum.
- Cross talk cased by other x-DSL, ISDN, or T1 lines in the cable bundle. This interference varies with frequency and can be severe.
- Impulse noise from lightning, appliances, and so on. It is typically of short duration and constant in frequency.

DMT accommodates line interference in a novel and sophisticated manner. It adapts to each subscriber line and assigns to each of the 256 discrete tones only that amount of data which it can accommodate to support reliable transmission. The adaptation is automatic and continuous and maximizes the available bit rate (Figs. 8.32 and 8.33).

There is a second, nonstandard, line code, which is in extensive use, called **carrierless amplitude phase (CAP)** modulation. It is a proprietary modulation scheme developed by AT&T and owned by Globespan Corporation. Its primary advantage is a significantly lower power requirement than DMT. It suffers, however, from a substantially lower noise margin than DMT. An STB is required to convert the digital signal to analog. Standard telephone

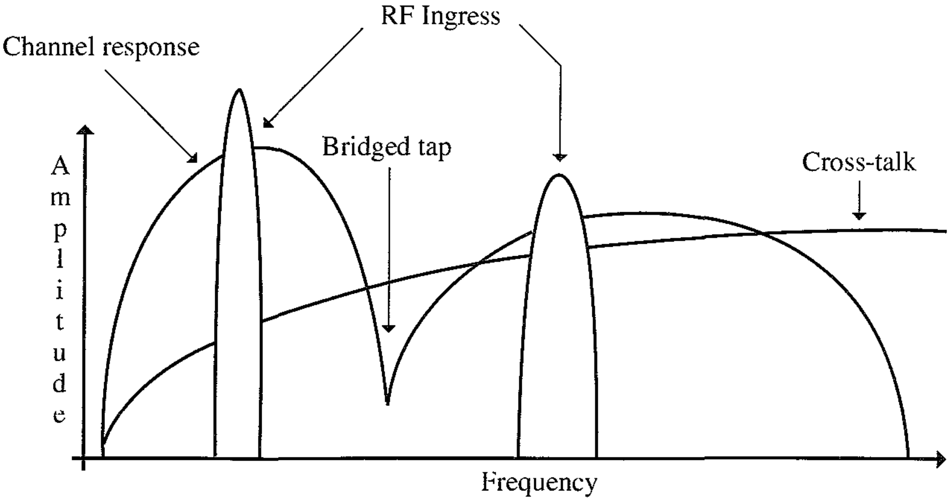


Figure 8.31 ADSL interference profile.

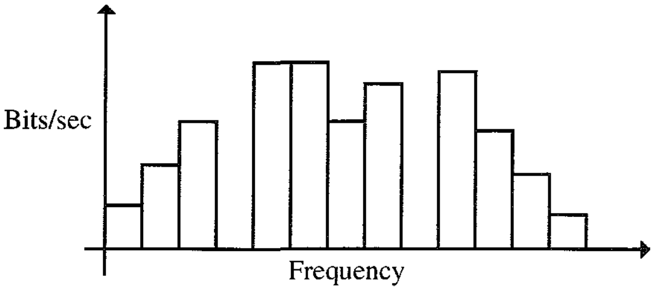


Figure 8.32 DMT modulation.

service is part of ADSL, as are other lower bandwidth upstream services (e.g., VOD/NVOD control, data communication, etc.); hence, the term "asymmetric."

Although ADSL is very attractive from the standpoint of cost—virtually every home has standard telephone wiring—it is much more bandwidth-limited than HFC. ADSL will carry approximately 6 Mbit/s from the pedestal to the home. For this reason, some telephone companies (e.g., Pacific Bell) have installed HFC systems in parallel with their copper pairs to take advantage of their ability to provide cable services under the Telecommunications Act of 1996.

The bandwidth limitation of ADSL prompted the development of the VDSL (see Ciaffi[6] for a complete description of xDSL). At a lower transmit power, standard 26-gauge copper telephone wire can carry 51 Mbit/s over 1.2 kft and 26 Mbit/s over 3 kft. ADSL would allow the transmission of a single 4-Mbit/s compressed video signal. For the subscriber with two or more television sets, this might be a severe limitation to a cable television type of service. VDSL, on the other hand, would provide 6 to 12 channels.

8.6 VIDEO COMPRESSION

A great deal of research and development has resulted in methods to reduce drastically the digital bandwidth required for video transmission. An uncompressed digital video signal takes roughly 150 Mbit/s of digital bandwidth. Early compression techniques compressed video

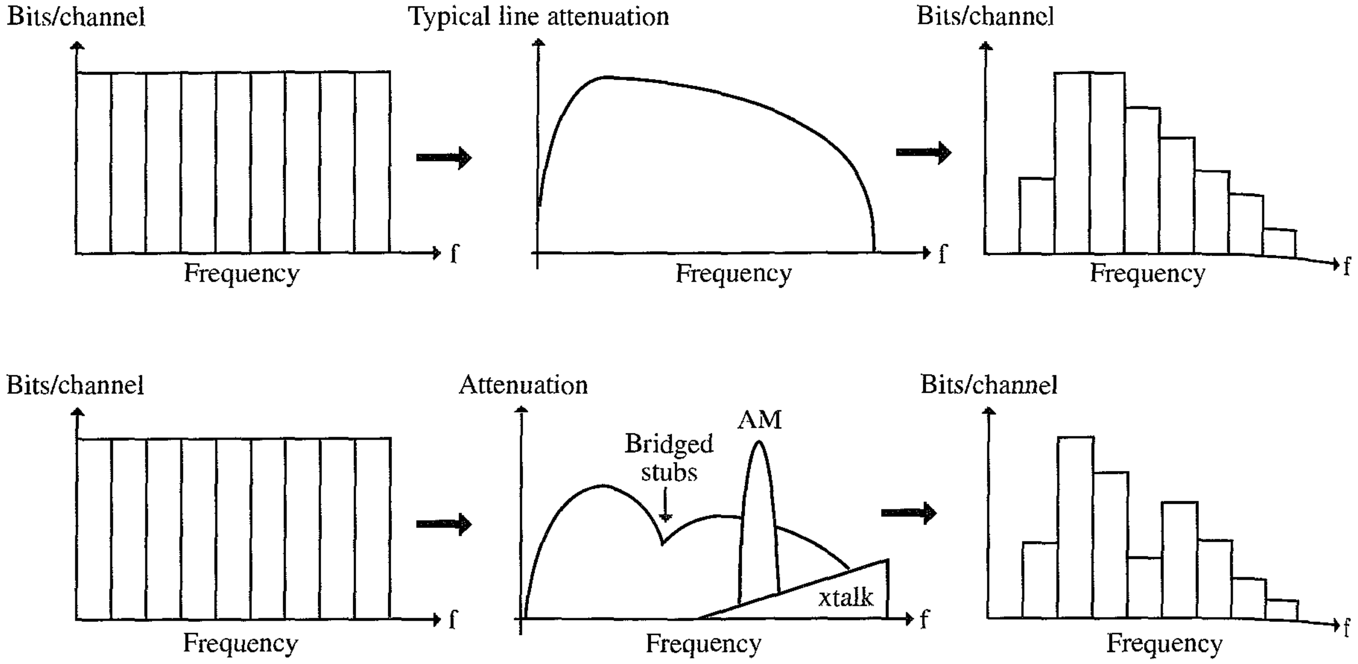


Figure 8.33 Adaptation characteristics.

signals to approximately 45 Mbit/s (DS3). For the emerging video delivery technologies of HFC, ADSL, HDTV, and so on, however, much greater compression was required. The Motion Picture Experts Group (MPEG) approached this problem and developed new compression techniques, which provide network or VCR quality video at much greater levels of compression. MPEG is a joint effort of the International Standards Organizations (ISO), the International Electotechnical Committee (IEC), and the American National Standards Institute (ANSI) X3L3 Committee.[6,7] MPEG has a very informative web site which provides extensive information on MPEG and JPEG technologies and standards (http://www.mpeg.org/index.html/).

The concept of digital video compression is based on the fact that, on the average, a relatively small number of pixels change from frame to frame. Hence, if only the changes are transmitted, the transmission bandwidth can be reduced significantly. Digitizing allows the noise-free recovery of the analog signal and improves the picture quality at the receiver. Compression reduces the bandwidth required for transmission and the amount of storage for a video program and, hence, expands channel capacity. A 2-hour digitized, but not compressed, NTSC video program would require roughly 100 gigabytes of storage.

There are two primary MPEG standards in use:

- *MPEG-1*: Used for VCR quality video and storage on CDs. MPEG-1 decoders are available on many PCs.
- *MPEG-2*: Provides VCR to HDTV quality transmission depending on data rate. It offers 50:1 compression of raw video.

(NTSC broadcast television in digital form requires 45–120 Mbit/s; MPEG-2 requires 1.5–15 Mbit/s. HDTV would require 800 Mbit/s uncompressed; MPEG-2 will transmit at 18 Mbit/s.)

There are two types of MPEG compression, which eliminate information that cannot be detected by the eye or ear:

1. Video
   - Spatial or intraframe compression, which forms a block identifier for a group of pixels having the same characteristics (color, intensity, etc.) for each frame. Only the block identifier is transmitted.
   - Temporal or interframe compression, which predicts interframe motion.
2. Audio, which uses a psychoacoustic model of masking effects.

The basis for video compression is to remove redundancy in the video signal stream. As an example of this approach, consider Fig. 8.34a and b. In Fig. 8.34a the car is in position *a* and in Fig. 8.34b it is in position *b*. Note that the background has remained essentially unchanged between the two figures. Figure 8.34c represents the information transmitted; that is, the change between the two frames. The car on the left represents the blocks of frame 1 which are replaced by background in frame 2. The car on the right represents the blocks of frame 1 which replace the background in frame 2.

Video compression starts with an encoder, which converts the analog video signal from the video camera to a digital format on a pixel-by-pixel basis. Each video frame is divided into 8 × 8 pixel blocks, which are analyzed by the encoder to determine which blocks must be transmitted, that is, which blocks have significant changes from frame to frame. This process takes place in two stages:

1. Motion estimation and compensation. This process involves a motion estimator identifying the areas or groups of blocks from a preceding frame that match corresponding areas in the current frame and transmitting the magnitude and direction of the displacement to a predictor in the decoder. The frame difference information is called the **residual.**
2. Transforming the residual on a block-by-block basis into more compact form.

The encoded residual signal is transformed into a more compact form by means of a **discrete cosine transform (DCT)** (see Haskel et al.,[6] sec. 6.5.2), which represents each pixel by a numerical value which is normalized for more efficient transmission. The DCT is of the form:

$$F(j,k) = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} f(n,m)\cos\left[\frac{(2n+1)j\pi}{2N}\right]\cos\left[\frac{(2m+1)k\pi}{2N}\right]$$

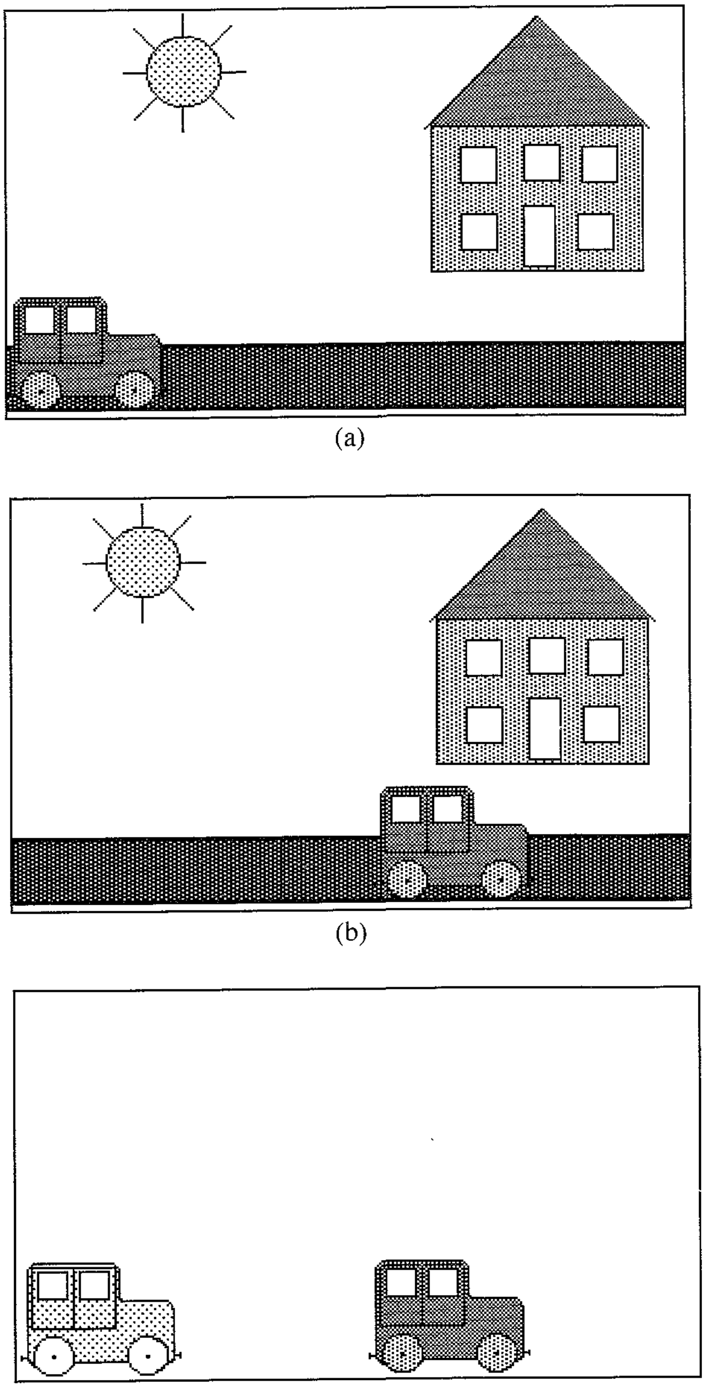
where $f(n,\ m)$ is the value assigned to the block in the $(n,\ m)$ position. The inverse transform is

$$f(n,m) = \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} F(j,k)\cos\left[\frac{(2n+1)j\pi}{2N}\right]\cos\left[\frac{(2m+1)k\pi}{2N}\right]$$

The DCT is typically multiplied, for an 8 × 8 block, by the expression $C(j)C(k)/4$, where

$$C(x) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } x = 0 \\ 1 & \text{otherwise} \end{cases}$$

Tables 8.9 and 8.10 depict the pixel block values before and after the DCT. One notes from Table 8.10 that there are relatively few meaningful elements, that is, elements with significant values relative to the values centered about the 0, 0 position. Because of this, most of the matrix values may be assumed to be zero, and, upon an inverse transformation, the



(a)

(b)

(c)

Figure 8.34 (a) Frame 1. (b) Frame 2. (c) Information transferred between frames 1 and 2

original values are quite accurately reproduced. This process greatly reduces the amount of data that must be transmitted, perhaps by a factor of 8 to 10 on the average. Note that the size of the transmitted residual may be that of an individual block or, at the other extreme, that of the entire picture.

The transformed matrix values of a block (Table 8.10) are normalized so that most of the values in the block matrix are less than 1. Then the resulting normalized matrix is quantized to obtain Table 8.11. Normalization is accomplished by a dynamic matrix of multiplicative

Table 8.9
8 × 8 Pixel block residual

	n							
	158	158	158	163	161	161	162	162
	157	157	157	162	163	161	162	162
	157	157	157	160	161	161	161	161
m	155	155	155	162	162	161	160	159
	159	159	159	160	160	162	161	159
	156	156	156	158	163	160	155	150
	156	156	156	159	156	153	151	144
	155	155	155	155	153	149	144	139

Table 8.10
Transformed 8 × 8 pixel block residual DCT coefficients

	j							
	1259.6	1.0	−12.1	5.2	2.1	1.7	−2.7	−1.3
	22.6	−17.5	6.2	−3.2	2.9	−0.1	−0.4	−1.2
	−10.9	9.3	−1.6	−1.5	0.2	0.9	−0.6	0.1
k	7.1	−1.9	−0.2	1.5	−0.9	−0.1	0.0	0.3
	−0.6	0.8	1.5	−1.6	−0.1	0.7	0.6	−1.3
	−1.8	−0.2	−1.6	−0.3	0.8	1.5	−1.0	−1.0
	−1.3	0.4	−0.3	1.5	−0.5	−1.7	1.1	0.8
	2.6	1.6	3.8	−1.8	−1.9	1.2	0.6	−0.4

values, which are applied element by element to the transformed matrix. The normalized matrix of Table 8.11 is the block information transmitted to the decoder. The denormalized matrix pictured in Table 8.12 and the reconstructed (inverse transformed) residual in Table 8.13 are determined by the decoder. The transformation proceeds in a zigzag pattern, as illustrated in Fig. 8.35.

MPEG approaches the motion estimation and compensation to remove temporal (frame-to-frame) redundancy in a unique way. MPEG uses three types of frames, the **intraframe** or **I-frame** (sometimes called the **independently coded** or **intracoded frame**), the **predicted (predictive)** or **P-frame**, and the **bidirectionally predictive frame** or **B-frame**. The P-frames are predicted from the I-frames. The B-frames are bidirectionally predicted from *either* past or future frames. An I-frame and one or more P-frames and B-frames make up the basic MPEG processing pattern, called a **group of pictures (GOP)**-(see Fig. 8.33). Most of the frames in

Table 8.11
Normalized and quantized residual DCT coefficients

	jn							
	21	0	−1	0	0	0	0	0
	2	−1	0	0	0	0	0	0
	−1	1	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0

Table 8.12
Denormalized DCT coefficients

	jn							
	1260	0	−12	0	0	0	0	0
	23	−18	0	0	0	0	0	0
	−11	10	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0

Table 8.13
Inverse DCT coefficients reconstructed residual

	n							
	158	158	158	163	161	161	162	162
	157	157	157	162	163	161	162	162
	157	157	157	160	161	161	161	161
m	155	155	155	162	162	161	160	159
	159	159	159	160	160	162	161	159
	156	156	156	158	163	160	155	150
	156	156	156	159	156	153	151	144
	155	155	155	155	153	149	144	139

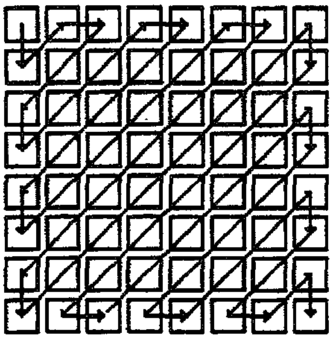


Figure 8.35 Zigzag DCT coefficient scanning pattern.

an MPEG compressed image are B-frames. The I-frame provides the initial reference for the frame differences to start the MPEG encoding process. Note that the bidirectional aspect of the procedure introduces a delay in the transmission of the frames since the GOP is transmitted as a unit and, hence, transmission must wait until the GOP is complete (Fig. 8.36). The detail of the procedure is beyond the scope of this text. (However, see, for example, section 7 "Source Compression."[8] In addition, one may find numerous references to MPEG compression and HDTV on the Internet.)

8.7 HIGH-DEFINITION TELEVISION (HDTV)

High-definition television (HDTV) is one of the **advanced television (ATV)** functions along with 525-line compressed video for **direct broadcast satellite** (DBS) or cable. The concept

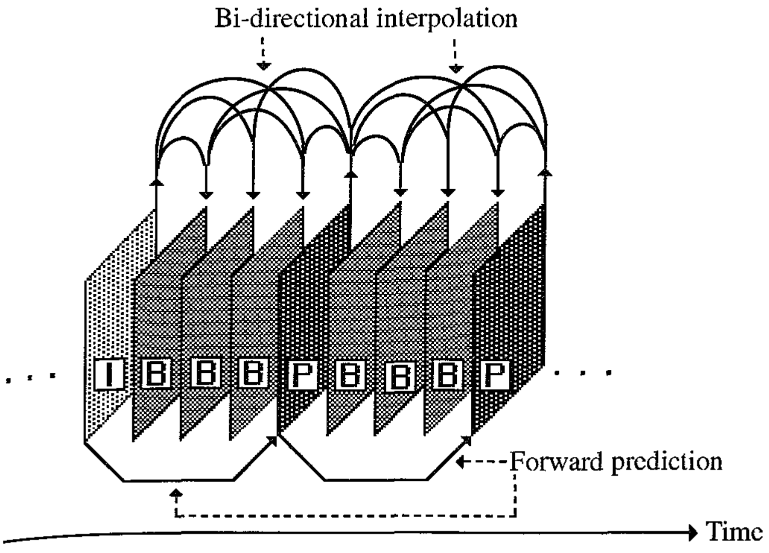


Figure 8.36 MPEG temporal frame structure.

of HDTV appeared in the late 1970s. Early development work was performed primarily in Japan based on an analog system. In the mid-1980s it became apparent that the bandwidth requirements of an analog system would be excessive, and work began on a digital system which could utilize the 6-MHz bandwidth of NTSC television. In the early 1990s seven digital systems were proposed, but testing of the seven systems indicated that none would be satisfactory. In 1993 the FCC suggested that the industry form a "Grand Alliance" (GA) to develop a common HDTV standard. Such a standard was developed and approved, in concept, by the FCC in August 1996.

The GA HDTV standard is based on a 16:9 aspect ratio (motion picture aspect ratio) rather than the 4:3 aspect ratio of NTSC television. HDTV will use MPEG-2 compression at 18 Mbit/s and a digital modulation format called 8-VSB (vestigial sideband), which uses an eight amplitude level symbol to represent 3 bits of information. Transmission is in 207-byte blocks, which include 20 parity bytes for forward error correction. The remaining 187-byte packet format is a subset of the MPEG-2 protocol and includes headers for timing, switching, and other transmission control.

The Advanced Television Systems Group, the successor to the Grand Alliance, has been developing standards and recommended practices for HDTV. These are found, along with a great deal of other information, on their web site: http://www.atsc.org/.

REFERENCES

1. B. P. Lathi and M. Wright, "Digital Hierarchy," in *The Communications Handbook*, J. D. Gibson, Ed., CRC Press, Boca Raton, FL, 1996, chap. 28.
2. W. Stallings, *Data and Computer Communications*, 3rd ed., Macmillan, New York, 1991, sec. 4-2.
3. J. Bellamy, *Digital Telephony*, 2nd ed., Wiley, New York, 1991, sec. 5.5.
4. H. J. Helgert, *Integrated Services Digital Networks*, Addison-Wesley, Reading, MA, 1991, chap 1.
5. J. M. Cioffi, "Digital Subscriber Lines," in *The Communications Handbook*, J. D. Gibson, Ed., CRC Press, Boca Raton, FL, 1996, chap. 34.
6. B. G. Haskel, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2*, Chapman and Hall, New York, 1996

7. J. L. Mitchell, W. B. Pennebaker, C.E Fogg and D. J. LeGall, *MPEG Video Compression Standard*, Chapman and Hall, New York, 1996.
8. "Source Compression," in *The Communications Handbook*, J. D. Gibson, Ed., CRC Press, Boca Raton, FL, 1996.

PROBLEMS

8.1-1 If a plesiochronous network operates from a cesium beam clock which is accurate to ±3 parts in 10^{12}, how long will it take for a DS3 signal transmitted between two networks to become out of sync if a 1/4-bit-length time error results in desynchronization?

8.1-2 For the bit stream **011100101001111011001** draw an AMI waveform (Fig. 8.1-2).

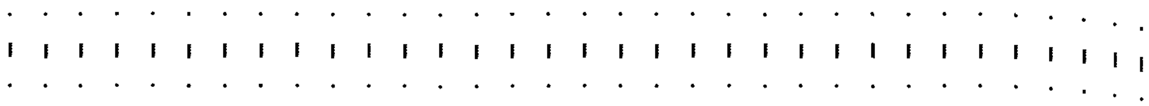
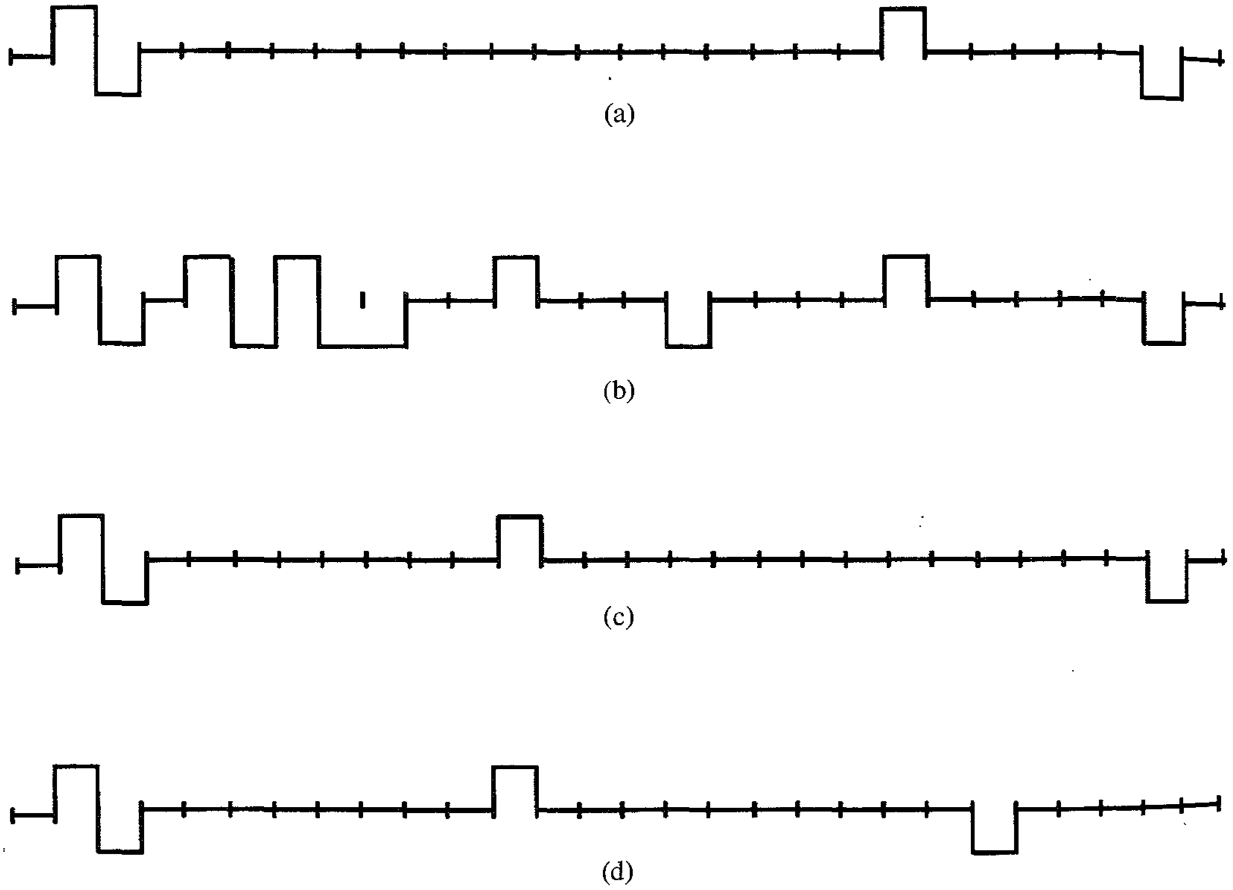


Figure P8.1-2

8.1-3 For the waveforms in Fig. P8.1-3, determine if each is a valid AMI format for a DS1 signal. If not, explain why not.



(a)

(b)

(c)

(d)

Figure P8.1-3

8.1-4 (**a**) You have received the following sequence of ESF framing pattern sequence bits:

. . . 00110010110010110 . . .

Is this a legitimate framing bit sequence in order to maintain synchronization between the T1 transmitter and receiver? If yes, why? If no, why not?

(**b**) The T1 AMI signal shown in Fig. P8.1-4 is received.
Is this an acceptable T1 signal?

(**i**) If yes, explain.

(**ii**) If no, explain why not (what, if any, DS1 standards are violated?) and draw on the figure the AMI waveform that would be transmitted by the DSU.

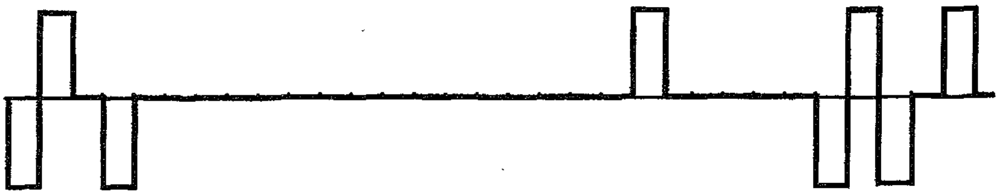


Figure P8.1-4

8.1-5 The signal **11010000000000000001** is received by the DSU in a T1 data stream which uses a B8ZS format. Draw the output of the DSU for this signal (Fig. P8.1-5). The first **1** is already drawn. Show the bit stream that is substituted by the DSU.

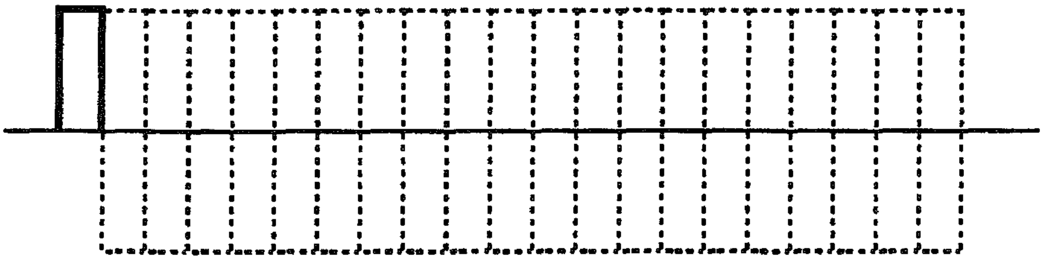


Figure P8.1-5

8.1-6 T1 synchronization at two distant locations is controlled by separate crystal-controlled oscillators which differ in frequency by 125 parts per million. If the terminal equipment does not maintain synchronization, in how many complete D4 superframes will the faster oscillator have generated (at most) one more time slot (8 bits) than the slower oscillator? (**a**) 5; (**b**) 10; (**c**) 15; (**d**) 20; (**e**) None of the above. If none, what is the number of D4 superframes before an extra time slot is generated?

8.1-7 Two plesiochronous digital networks, A and B, utilize cesium beam clocks accurate to 3 parts in 10^{13}. The networks are operated by independent long-distance companies and are synchronized to each other by means of a UTC signal.

(**a**) If a company leases a T1 line with D4 framing, which is terminated at one end in network A and at the other end in network B, how often must the networks be resynchronized to each other to avoid a framing bit error in the customer's T1 signal in the worst case? *Hint*: You may assume that a framing bit error occurs when the two networks are out of synchronization by $\geq 1/2$ of a T1 "bit time."

(**b**) UTC operates via GPS satellites, which are approximately 23,000 miles above the earth. How long, in terms of T1 bits, will a correction signal take to be transmitted to the network switches?

9 SOME RECENT DEVELOPMENTS AND MISCELLANEOUS TOPICS

In this chapter, we discuss some recent developments in communication technology along with various communication media.

9.1 CELLULAR TELEPHONE (MOBILE RADIO) SYSTEM

The wireless revolution of cellular phones began just over 10 years ago. But during this short time the cellular phone has changed from a status symbol to a necessity. This was helped by the dramatic cost reduction from a high of about \$2000 to \$100.

The cellular phone service area is divided into smaller geographical areas called **cells** (Fig. 9.1). Each cell has a **base station** with a tower, which receives and transmits phone signals to mobile users. All the base stations are connected by telephone lines to the **mobile telephone switching office (MTSO)**, which in turn is connected to the telephone central office by phone lines. A caller communicates via radio channel to a cell-site base station, which sends the signal to the MTSO. If the called party is land based, the MTSO sends the signal through the central office like any other telephone call. If the called party is mobile, MTSO sends the signal to the base station of the cell where the called party is located. The base station transmits the signal to the called party using the available radio channel in the cell. As the caller moves from one cell to another, the MTSO automatically switches the user to an available channel in the new cell while the call is in progress.

Each cellular phone has a manufacturer's serial number and the phone number assigned by the phone company. These numbers are automatically transmitted to MTSO during the initialization of the call. The MTSO, after authenticating the numbers, assigns to the caller two available frequencies (radio channels), one for transmission to and the other for receiving from the base station. When the call terminates, the radio channels become available for another user. The MTSO continuously monitors the signal strength of a phone call, and the signal attenuation beyond some point is viewed as an indication of the caller moving from a previous cell to the next cell. MTSO then searches for a neighboring cell, where the signal strength from

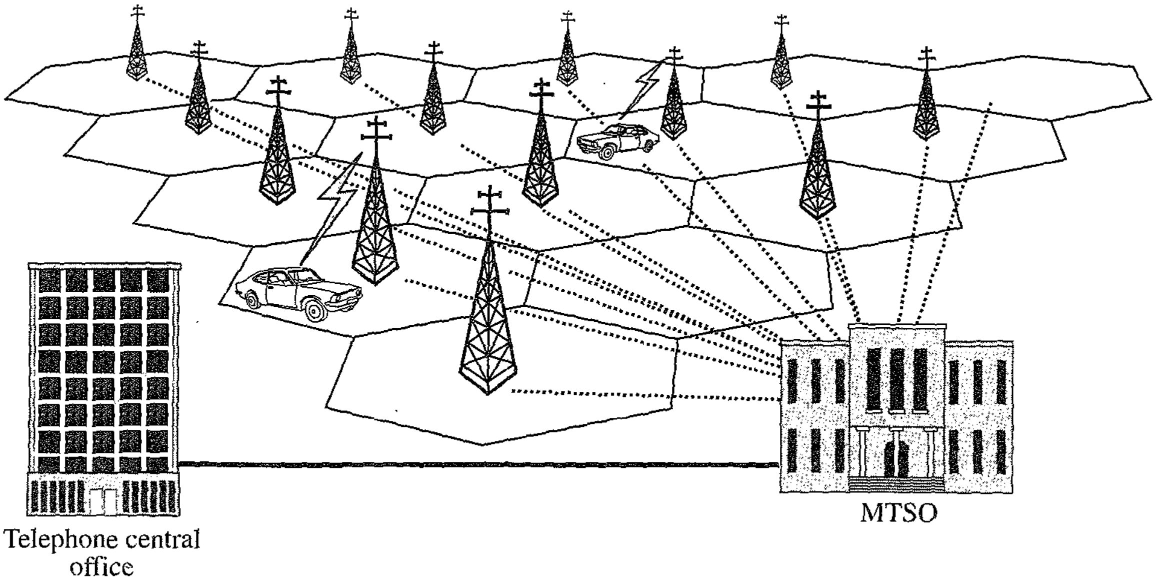


Figure 9.1 Cellular telephone system.

the caller is stronger and then automatically switches the caller to the next base station. The switching is so rapid that users do not notice it.

Before cellular communication, high-power transmitters were used so that each radio channel covered the entire city. Hence, a channel could be used only by one user in the city. This posed serious limitations on the number of channels available, which restricted the number of phones in a given area. This limitation is overcome in the cellular scheme by reusing the same frequencies in all the cells except those immediately adjacent. This is possible because the transmitted powers are kept sufficiently small so that the signals from one cell do not propagate beyond the immediately adjacent cells. We can accommodate any number of users by increasing the number of cells as we reduce the cell size and the power levels correspondingly.

The analog cellular schemes use a 3-kHz audio signal to frequency-modulate a carrier with transmission bandwidth 30 kHz ($\beta = 4$). This wide-band FM signal results in a high SNR. But FM does not utilize the exchange of bandwidth for SNR as efficiently as PCM. Hence, the next generation cellular systems, which are just entering the market, are digital (see Sec. 9.2.3). The digital schemes have much higher quality reception and no cross talk (hearing other people's conversation).

A combination of satellite communication and cellular phones is also moving closer to reality. Motorola has undertaken a global satellite based cellular network (iridium telecommunication system). The network would relay telephone calls via 66 small satellites launched into low earth orbit (LEO) at altitudes of 900 to 10,000 km. Voice transmission via LEO is expected to suffer less from the echo found on geostationary satellite links, because the round-trip signal time to LEO satellites is less. Also, because of the satellites' proximity to earth, user handsets will require less power for transmitting and receiving signals, and can also be smaller, lighter, and cheaper. The drawback of LEO is that we need many more satellites for coverage of the entire earth (66 in all). The current plan calls for launching all the satellites over two years beginning in 1996. Commercial services will begin in 1998.

9.2 SPREAD SPECTRUM SYSTEMS

The interest in wireless communications has increased dramatically in the last five years, as seen by the growth of cellular telephony, radiopaging, personal communication systems, and indoor applications such as wireless local-area networks. This increased interest has brought more focus on the problems unique to the wireless environment, including capacity limits due to spectrum availability, propagation effects such as multipath propagation, and the need for asynchronous access. One possible method of addressing the aforementioned problems is the use of spread spectrum communications. Spread spectrum promises several benefits, such as higher capacity and the ability to resist multipath propagation.

Historically spread spectrum was developed for secure communication and military uses. Spread spectrum signals have the following characteristics:

1. They are difficult to intercept for an unauthorized person.
2. They are easily hidden. For an unauthorized person, it is difficult to even detect their presence in many cases.
3. They are resistant to jamming.
4. They provide a measure of immunity to distortion due to multipath propagation.
5. They have an (asynchronous) multiple-access capability.

The spread spectrum refers to any system that satisfies the following conditions:

1. The spread spectrum may be viewed as a kind of modulation scheme in which the modulated (spread spectrum) signal bandwidth is much greater than the message (baseband) signal bandwidth. Thus, spread spectrum is a wideband scheme, such as, wide-band angle modulation.
2. The spectral spreading is performed by a code that is independent of the message signal. This same code is also used at the receiver to despread the received signal in order to recover the message signal (from the spread spectrum signal). In secure communication, this code is known only to the person(s) for whom the message is intended. Wide-band angle modulation is not a spread spectrum scheme because its spectral spreading is done, not by an independent code, but by the message signal itself.

The spread spectrum scheme increases the bandwidth of the message signal by a factor N, called the **processing gain**. If the message signal bandwidth is B Hz and the corresponding spread spectrum signal bandwidth is B_{ss} Hz, then

$$\text{Processing gain } N = \frac{B_{ss}}{B}$$

Although we use much higher bandwidth for a spread spectrum signal, we can also multiplex large numbers of such signals over the same band by assigning a different code to each signal. The codes are so chosen as to achieve near orthogonality of the waveforms. Recall that the two carriers used in QAM are also orthogonal. This orthogonality, if strict, would allow multiple users to coexist in a given frequency range without mutual interference, providing multiple access through what is known as **code-division multiple access (CDMA)**. This same principle also makes spread spectrum less vulnerable to intentional or unintentional interference.

There are several forms of spread spectrum. We discuss here the two primary techniques of spread spectrum signals, the **direct sequence spread spectrum (DS/SS)** and the **frequency hopping spread spectrum (FH/SS)**. In addition to DS/SS and FH/SS, hybrid formats of the two schemes are also used. Other forms such as time hopping and chirp modulation, although studied, are not common and will not be discussed here.

9.2.1 Direct Sequence Spread Spectrum (DS/SS)

In spread spectrum systems, the signal spreading code is the so-called **pseudonoise (PN)** sequence, which is generally periodic and consists of a periodic coded sequence of **1**'s and **0**'s with certain autocorrelation properties. These signals are pseudorandom inasmuch as they appear to be unpredictable to an outsider, though they can be generated by deterministic means by the person for whom they are intended. The polar signal $c(t)$ representing this binary sequence (Fig. 9.2) is the pseudorandom carrier that is used to multiply the message signal $m(t)$ (which is also a polar binary signal) to obtain the DS/SS signal $y(t)$:

$$y(t) = m(t)c(t)$$

Figure 9.3 shows the generator of the DS/SS signal $y(t)$. The spectral spreading signal $c(t)$ is a pseudorandom signal inasmuch as it appears to be unpredictable, though it can be generated by deterministic means (hence, pseudorandom). The bit rate of $c(t)$ is chosen to be much higher than the bit rate of $m(t)$. The basic pulse in $c(t)$ is called the **chip**, and the bit rate of $c(t)$ is known as the **chip rate**. If the chip width is T_c, then the chip rate is $R_c = 1/T_c$.

The autocorrelation function $\psi_c(\tau)$ of $c(t)$ is very narrow, that is, it is concentrated around $\tau \le |T_c|$ and is small for $\tau > |T_c|$.[*] Furthermore, in the multiuser CDMA setting, the crosscorrelation $\psi_{c_1c_2}(\tau)$ between any two codes is very small in order to have negligible interference between various multiplexed signals. Such codes are generated by some form

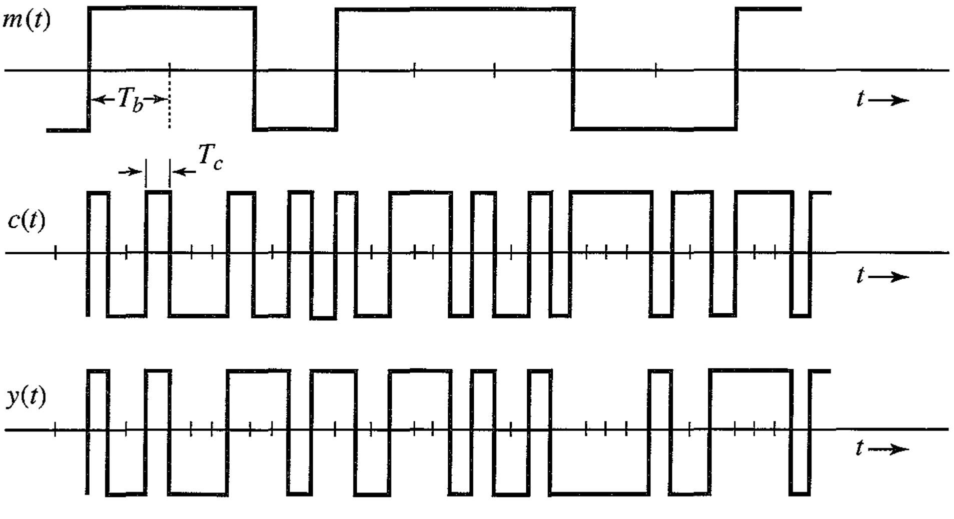


Figure 9.2 Signals at the spread spectrum generator.

---

[*] Because $c(t)$ is periodic with a period of L bits, its autocorrelation function is also periodic with the same period (L bits).

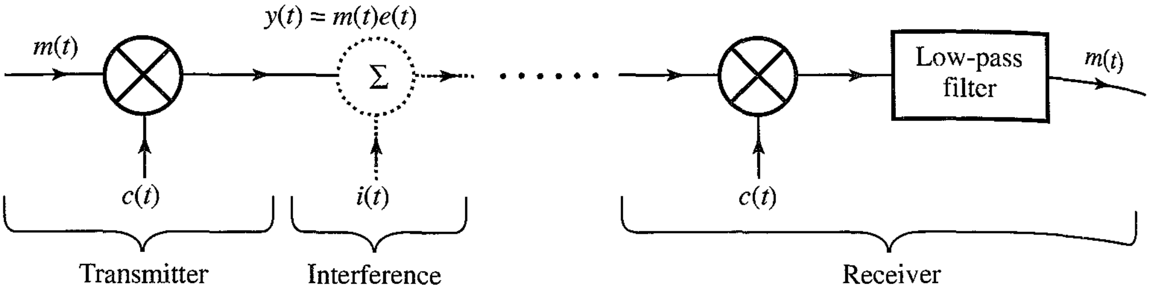


Figure 9.3 A direct sequence spread spectrum system.

of shift-register networks with output feedback capable of producing a sequence with long periods and, preferably, low susceptibility to structural identification by an eavesdropper (see Fig. 13.17).

Detection

At the receiver, we generate a synchronous version of the pseudorandom sequence $c(t)$ used at the transmitter. The received DS/SS signal $y(t)$, when multiplied by $c(t)$ (Fig. 9.3), yields the desired signal $m(t)$ because

$$y(t)c(t) = m(t)c^2(t) = m(t) \qquad \text{because } c^2(t) = 1$$

Thus, the process of detection (despreading) is identical to the process of spectral spreading. Recall that for DSB-SC, we have a similar situation in that the modulation and demodulation processes are identical (except for the output filter). Note that a low-pass filter is basically an integrator. Hence, the operation at the receiver in Fig. 9.3 is performing correlation between the incoming signal $m(t)c(t)$ and the locally generated signal $c(t)$. The discussion here is kept at a relatively simple level. A rigorous approach to this topic, which requires an understanding of optimum receivers, is given in Chapter 13.

DS/SS Coherent PSK: The preceding discussion relates to the baseband transmission. To use this scheme over bandpass channels, we further DSB-SC modulate the baseband DS/SS signal to obtain the corresponding PSK version $y(t)\cos\omega_c t$. At the receiver this incoming signal is synchronously demodulated to obtain the baseband DS/SS signal $y(t)$, which is then detected, as shown in Fig. 9.3.

Signal Spectra

In Chapter 4, we showed that the minimum theoretical bandwidth of a binary signal at a rate R_b bit/s is $R_b/2$ Hz (two pieces of information per second per hertz). We also saw that the bandwidth of practical binary schemes is proportional to the bit rate. Figure 9.2 shows that the bit rate of the spread spectrum signal $y(t)$ is the same as the chip rate [the rate of $c(t)$]. If the chip rate is R_c bit/s, and the message symbol rate is R_b, then the bandwidth of $y(t)$ is R_c/R_b times the bandwidth of $m(t)$. Thus, the processing gain is

$$N = \frac{R_c}{R_b} = \frac{T_b}{T_c}$$

The PSD of a polar binary signal using a half-width pulse is shown in Fig. 7.5. For a full-width pulse, $p(t) = \text{rect}\,(t/T_b)$ and $P(\omega) = T_b \,\text{sinc}\,(\omega T_b/2)$. Hence, from Eq. (7.12),

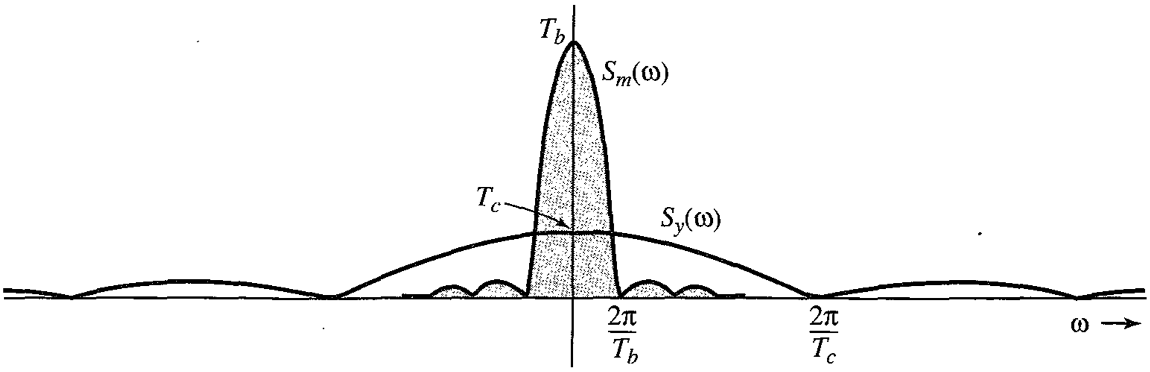


Figure 9.4 PSDs of the input and the output signals of a DS/SS system.

$$S_m(\omega) = T_b \operatorname{sinc}^2\left(\frac{\omega T_b}{2}\right)$$

The chip width is T_c. Hence, the PSD of $y(t)$ is

$$S_y(\omega) = T_c \operatorname{sinc}^2\left(\frac{\omega T_c}{2}\right)$$

The PSDs of $m(t)$ and $y(t)$ are shown in Fig. 9.4. Because the PSD bandwidth is directly proportional to the bit rate (or inversely proportional to the pulse width), the PSD of $y(t)$ is wider than the PSD of $m(t)$ by a factor N. But their powers are identical. This means the PSD of $y(t)$ is weaker than that of $m(t)$ by a factor N (the processing gain). It also means that the spreading process reduces the PSD of a signal by a factor N, the processing gain.

How Does DS/SS Realize Its Special Features?

Secure Communication and Jamming Resistance: The secure communication feature of the DS/SS is due to the fact that this signal can be detected only by authorized person(s) who know the pseudorandom code used at the transmitter. Also because the DS/SS signal spectrum is spread over a very wide band, the signal PSD is very small, which makes it easier to hide the signal within the noise floor. Moreover, because of the distribution of the signal power over a wide band, spread spectrum signals are difficult to jam. Jamming is effective only if the signal to be jammed occupies a smaller band. We can explain the ineffectiveness of jamming by referring to Fig. 9.3. The interfering signal $i(t)$ is the jamming signal that is added to the DS/SS signal $y(t)$ along the communication path, as shown in Fig. 9.3. The received signal $y(t) + i(t)$ is now multiplied by $c(t)$ at the detector to yield

$$[y(t) + i(t)]c(t) = [m(t)c(t) + i(t)]c(t) = m(t)c^2(t) + i(t)c(t) = m(t) + i(t)c(t)$$

Observe that the detector despreads the signal $y(t)$ to yield $m(t)$. On the other hand, the jamming signal $i(t)$ is spread to yield $i(t)c(t)$. In other words, the PSD of the jamming signal, because of spectral spreading, decreases by a factor N, the processing gain, as explained earlier. However, the PSD of the desired signal $m(t)$ becomes stronger because of despreading (Fig. 9.4). Using a low-pass filter, we can recover $m(t)$ with only a small fraction of the power from $i(t)$. Understandably, spread spectrum found its original application in military communication systems.

It may appear that channel noise, being an interference, will also suffer a power reduction by a factor N at the receiver. Unfortunately, this does not happen because the noise is a broadband signal to begin with, and it is impossible to spread it further. The spectral spreading can be realized only if the interference bandwidth is on the order of the baseband signal.

Multiple Access—Several Users on the Same Band: In a spread spectrum system, several users can utilize the same band. We can view the multiple-access problem much the same way we view the case of jamming. If individual users have independent, uncorrelated spreading codes, undesired signals (cochannel interference) will not be despread in the receiving process. Ideally, for the case of two users, this provides a signal-to-interference ratio of N, the processing gain.

Advantages of CDMA over TDMA and FDMA: The DS/SS allows greater capacity by allowing multiple-access communication. **Frequency-division multiple access (FDMA)** and **time-division multiple access (TDMA)** have a fixed number of frequency or time slots available to the user. In addition, the frequencies cannot be reused unless geographical areas using the same frequencies are separated by sufficient distance to avoid cochannel interference. In spread spectrum, several users can occupy the same frequency spectrum simultaneously, and frequency bands can be reused without regard to the separation distance of the users. This is because all users have unique spreading codes, which are ideally mutually orthogonal, giving rise to the term code-division multiple access (CDMA).

An additional implication of the low crosscorrelation between users is that, unlike TDMA or FDMA, there is no hard limit on the number of users allowed in the system at one time. As the number of users increases, the signal quality of all users degrades gracefully, placing a soft limit on the number of users.

Resistance to Multipath Fading: To understand the resistance of DS/SS to multipath fading, we note that the signal received from any undesired path is a delayed version of the DS/SS signal. But the DS/SS signal has a property of low autocorrelation (small similarity) with its delayed version, especially if the delay is of more than one chip duration. Hence, the delayed signal, looking more like an interfering signal, will not be despread by $c(t)$. This effectively minimizes the effect of the multipath signals.

What is more interesting is that DS/SS cannot only mitigate but may also exploit the multipath propagation effect. This is accomplished by a **Rake** receiver. This receiver is so designed as to coherently combine the energy from several multipath components, which increases the received signal power and thus provides a form of diversity reception. The Rake receiver consists of a bank of correlation receivers, with each individual receiver correlating with a different arriving multipath component. By adjusting the delays, the individual multipath components can be made to add coherently rather than destructively.

Near-Far Problem: The DS/SS form of spread spectrum has the best performance in terms of jamming rejection and multipath immunity. But it does suffer from the **near-far** problem. The discussion so far assumes that the signals from all users are received with the same signal power. When this is not true, we may encounter the near-far problem. The despreading of a desired signal increases its strength N-fold compared to the residual noise level due to other unwanted signals. However, if an unwanted signal strength is strong due to the proximity of its transmitter to the receiver, and the strength of the desired signal is weak due to the remoteness of its transmitter from the receiver, the undesired signal may still drown out the desired signal. This often necessitates adaptive power control techniques to overcome the near-far problem. In addition, the interference from multiple users, although small individually, adds up to degrade the system performance. If all the codes were strictly orthogonal, this problem would not arise. Unfortunately, it is difficult to find a large number of codes that are strictly orthogonal, and we have to use many codes that are nearly orthogonal.

9.2.2 Frequency Hopping Spread Spectrum (FH/SS)

In frequency shift keying (FSK) discussed in Sec. 7.8 (Fig. 7.28), we transmit **0** and **1** using sinusoid pulses of frequencies ω_0 and ω_1, respectively. In a frequency hopping spread spectrum (FH/SS) scheme (Fig. 9.5), we generate such an FSK signal and then shift the frequency of the FSK signal by an amount determined by a PN code. The frequency shifting is performed by a frequency mixer (converter), as shown in Fig. 4.7. The frequency mixer is discussed in Example 4.2. If the FSK frequency is ω_i and the synthesizer frequency is ω_h, then the mixer yields outputs of frequencies $\omega_h + \omega_i$ (sum) and $\omega_h - \omega_i$ (difference). The bandpass filter selects the sum frequencies and suppresses the difference frequencies. The data is now transmitted by sinusoidal pulses, whose frequencies hop over a wide range of frequencies according to a PN code. The carrier frequencies used in these hops are separated so that the resulting signal spectra in various hops are nonoverlapping. The bandwidth of the FH/SS signal is, thus, many times that of the ordinary FSK signal. At the receiver, a frequency mixer with frequencies controlled by an identical PN code synchronized with the received signal shifts the frequencies back to original FSK frequencies. The resulting FSK signal is then demodulated (noncoherently), as discussed in Sec. 7.8.

The frequency hopping rate R_h may be less than or greater than the symbol rate R_b. If $R_h \leq R_b$, the FH is a **slow hopping (SFH)** scheme, and if $R_h > R_b$, the FH is a **fast hopping (FFH)** scheme. In an SFH, several symbols are transmitted in each frequency hop. On the other hand, in FFH, the carrier frequency will change or hop several times during the transmission of one symbol. FFH is used to defeat the jammer's tactic of detecting and then jamming the detected frequency. FFH is so fast that by the time a jammer detects and then switches to the target frequency, the target frequency has already changed.

There are some unique advantages of FH/SS. Because FH/SS can produce signals of much wider bandwidth compared to DS/SS, it can achieve much higher processing gains than

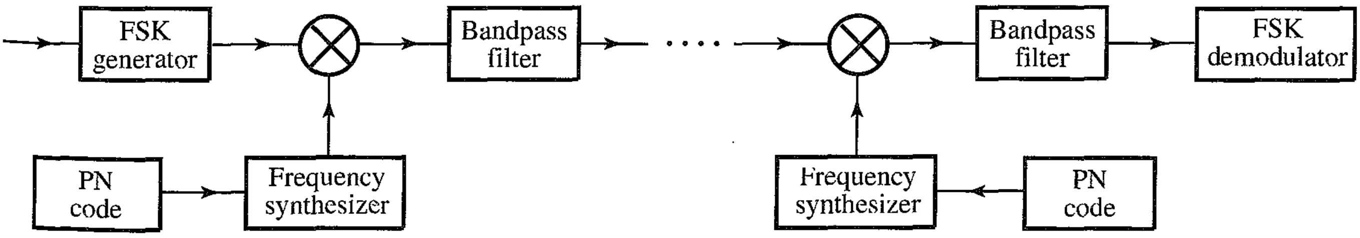


Figure 9.5 Frequency hopping spread spectrum system.

DS/SS. Also, FH/SS is less susceptible to near-far problems than DS/SS. This is because in FH/SS, (cochannel) signals are generally not utilizing the same frequency simultaneously. Hence, the relative power levels of (cochannel) signals are not as critical as in DS/SS. However, the FH/SS scheme does not have the same degree of jamming resistance as DS/SS. In FH/SS, the jamming resistance is achieved by randomly hopping the frequency and, thus, avoiding the jammer frequency. The jammer may also be constantly hunting for the target frequency, and occasionally collisions do occur. In such a case, data will be lost, especially in SFH, even when burst error correcting codes are used. In addition, SFH signals have only a limited multipath immunity. If the carrier hops to a frequency where the multipath transmission null resides, that part of the data will be lost. For FFH, several frequency hops occur during each symbol, and even if transmission null occurs at some of these frequencies, the data can still be recovered.

Although DS/SS is currently more important, some commercial applications of FH/SS are emerging. FH/SS is used in some spread spectrum devices that provide low-rate data communication in the unlicensed **industrial, scientific, and medical (ISM)** frequency band from 902 to 928 MHz, and both the **cellular and digital packet data (CDPD)** and **Global system for mobile communications (GSM)** standards have options for incorporating frequency hopping.

Spread spectrum is now finding widespread civilian and commercial applications in such areas as cellular telephones, personal communications, and position location. For example, DS/SS is used in Electronic Industries Association's Interim Standard IS-95[1] for digital cellular telephones, as well as a wide range of position location systems such as the global position location system to be discussed in the next section, and other vehicle location and messaging systems. In addition, there are a host of systems that are operating in the industrial, scientific, and medical band in which the FCC waives its usual licensing requirements for spread spectrum systems.

9.2.3 Applications of Spread Spectrum

Recently, commercial applications have begun to emerge for spread spectrum, including cellular telephones, personal communications, and position location. Here we discuss two applications of CDMA technology that illustrate the benefits of spread spectrum. This explanation can help understand CDMA's usefulness in terms of the parameters defined previously.

Cellular Telephony

Although spread spectrum is inherently well suited for multipath environments and affords a number of advantages in the areas of networking and hand-off, the key characteristic underlying the current interest in CDMA for wireless cellular systems is the potential for increased spectral efficiency. The capacity improvement comes from two key issues. First, the use of CDMA allows improved frequency reuse. Narrow-band systems cannot use the same transmission frequency in adjacent cells because of the potential for interference. CDMA has inherent resistance to interference. Although users from adjacent cells will contribute to the total interference level, their contribution will be significantly less than the interference from the same cell users. The result is that frequency reuse efficiency increases by a factor of 4 to 6. In addition, when used to transmit voice signals, CDMA systems may exploit the fact that voice activity typically lies at somewhat less than 40%, thus reducing the amount of interference to 40% of its original value. As a result, CDMA systems may use statistical multiplexing of voice signals (or turning off the transmitter when not needed) to further increase capacity.

IS-95, the wide-band digital standard adopted by the Electronic Industries Association,[1] employs DS/SS in order to provide dramatically higher capacity when compared to existing analog systems and possibly more than proposed TDMA systems. The system uses a 1.2288-MHz code spreading sequence on top of a variable data rate that ranges from 1200 to 9600 bit/s. Two levels of spreading codes are employed. A long code with a period of over one century is used to create a large number of potential spreading codes, partly in reaction to the concern that the number of unique codes in more common sets such as m-sequences or Gold codes was too small. A second, short code is also used for spreading to enable convenient synchronization. Identical data is transmitted over the in-phase (I) and quadrature (Q) channels using different short codes. Simultaneous transmission over the I and Q channels allows extremely sharp pulse shaping.

There are important differences between the forward and reverse links in the IS-95 standard. The performance of the reverse link is of greater concern for two reasons. First, as discussed in Sec. 13.4.4, the reverse link is subject to near-far effects. Second, since all transmissions on the forward link originate at the same base station, it is possible to generate synchronous signals with orthogonal spreading codes which result in zero crosscorrelation in Eq. (13.40b). For this reason, more powerful error correction is employed on the reverse link.*

As mentioned earlier, the near-far problem needs to be addressed when spread spectrum is utilized in mobile communications. To combat this problem IS-95 uses power control. On the forward link there is a subchannel for power control purposes. Every 1.25 ms the base station receiver estimates the signal strength of the mobile unit. If it is too high, the base transmits a **1** on the subchannel. If it is too low, it transmits a **0**. In this way, the mobile station adjusts its power every 1.25 ms as necessary so as to reduce interference to other users.

The QCELP (Qualcomm code-excited linear prediction) algorithm is used for voice encoding. Since the voice coder exploits gaps and pauses in speech, the data rate is variable. In order to keep the symbol rate constant, whenever the bit rate falls below the peak bit rate of 9600 kbit/s, repetition is used to fill the gaps. For example, if the output of the voice coder (and subsequently the convolutional coder) falls to 2400 bit/s, the output is repeated three times before it is sent to the interleaver. IS-95 takes advantage of this repetition time by reducing the output power during three out of the four identical symbols by at least 20 dB. In this way, the multiple-access interference is reduced. This voice activity gating reduces interference and increases overall capacity.

Global Positioning System (GPS)

A recent popular application that uses the spread spectrum is the **global positioning system (GPS)**, which found its original application in the military. GPS however, is now finding several uses in civilian life such as ship and aircraft navigation, surveying, and geological studies. The potential applications of GPS are so vast that it has been called (with some exaggeration) the next utility (similar to gas, water, and electricity). GPS allows a person to determine the time and the person's precise location (lattitude, longitude, and altitude) anywhere on earth within the accuracy of inches. The person can also find the velocity with which he or she is moving.

---

* Data is encoded using a rate 1/2 constraint length 9 convolutional code followed by an interleaver on the forward channel, while a rate 1/3 constraint length 9 convolutional code followed by an interleaver is used on the reverse link. Interleaving is utilized to avoid large burst errors, which can be very detrimental to convolutional codes. (See Chapter 16 for convolutional and burst-error correcting codes.)

GPS receivers are fast becoming small enough and cheap enough to be carried by just about everyone, including hikers and boaters. By the end of this decade, everyone will be touched by GPS technology. Some obvious applications: Delivery vehicles will be able to pinpoint destination. Emergency vehicles will be more prompt. Shipping companies can continuously monitor the status and locations of their trucks, which helps remedy the problem of cargo theft and truck hijacking. International Cargo Management System of San Antonio puts solar-powered transmitters inside cargo containers. Information on their location is continuously monitored at headquarters through communication satellites. Whenever a container door is opened, the system transmits that fact, alerting the company to hijacking on the spot. Cars will have electronic maps that will instantly show us the way to any destination. Hertz and Avis both rent out cars with GPS maps. Oldsmobile now offers GPS as a $2495 option, and other car makers plan to follow suit. The motorist keys in a destination and the system shows how to get there. With GPS-assisted navigation systems airlines can fly more direct routes, saving time for passengers and fuel cost for airlines. It is also the best (and cheapest) way to design a fool proof air collision avoidance system. Right now work is progressing on zero-visibility landing systems. Other interesting applications are surveying and geological studies. Recently, Boeing replaced its expensive inertial navigation system in its jumbo jets with GPS as a sole means of navigation. In the 1991 Gulf War, GPS was already providing pinpoint targeting for bombing and artillary fire, and U.S. troops used hand-held GPS receivers to guide them across the trackless desert with phenomenal accuracy. A GPS receiver was also responsible for the 1995 rescue of captain O'Grady in Bosnia. Dogged by Serb soldiers for 6 days, O'Grady used his GPS receiver to get his coordinates to a Marine Corps search-and-rescue team. Then the rescue helicopters used GPS receivers to navigate through the mountain fog and find O'Grady. While the pentagon spends hundreds of billions of dollars for the satellite system used for GPS, the use of the system is available to all, free of charge.

How GPS Works: A GPS receiver operates by measuring its distance from a group of satellites in space which are acting as precise reference points. The GPS system consists of 24 satellites, so there will always be more than four visible from anywhere on earth. The 24 satellites are at a height of 22,200 km located in six orbital planes. Each satellite circles the earth in 12 hours. The satellites are constantly monitored by the Department of Defense, which knows their exact locations and speeds at every moment. This information is relayed back to the satellites. All the satellites have atomic clocks of unbelievable precision on board and are synchronized so that they are generating the same PN code at the same time. The satellites are continuously transmitting this PN code and the information about their locations and time. A GPS receiver on the ground is also generating the same PN code, although not in synchronism with that of the satellites. This is because of the necessity to make GPS receivers inexpensive. Hence, the timing of the PN code generated by the receiver will be off by an amount of α seconds (timing bias) from that of the PN code of the satellites.

To begin with, let us assume that the timing bias $\alpha = 0$. By measuring the time delay between its own PN code and that received from one satellite, the receiver can compute its distance d from that satellite. This information places the receiver anywhere on a sphere of radius d centered at the satellite location (which is known), as shown in Fig. 9.6a. Simultaneous measurements from three satellites place the receiver on the three spheres centered at the

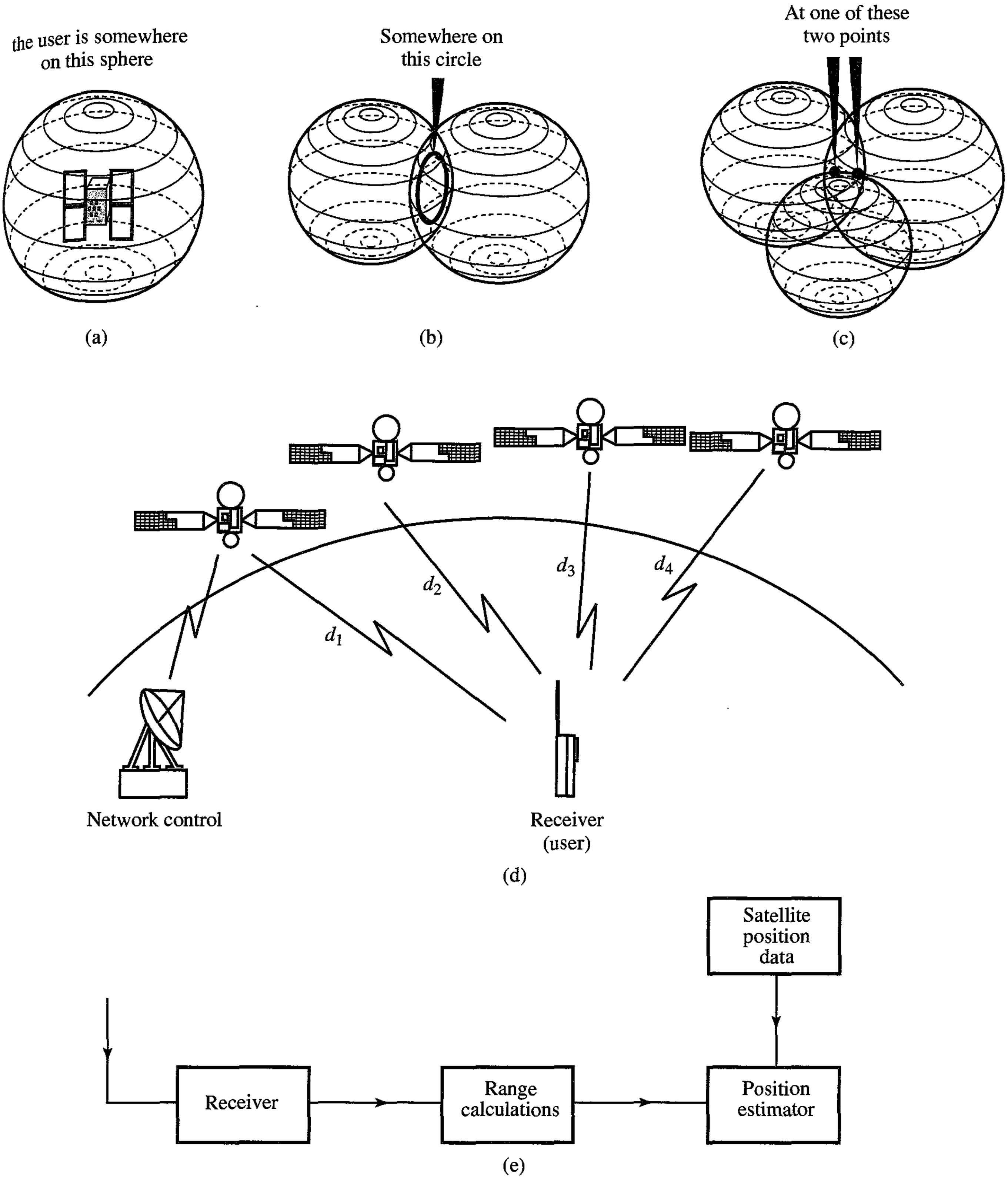

the user is somewhere on this sphere
Somewhere on this circle
At one of these two points
(a)
(b)
(c)
d_1
d_2
d_3
d_4
Network control
Receiver (user)
(d)
Satellite position data
Receiver
Range calculations
Position estimator
(e)

Figure 9.6 (a) Receiver location from one satellite measurement. (b) Location narrowed down by two satellite measurements. (c) Location narrowed down by three satellite measurements. (d) Practical global positioning system using four satellites. (e) Block diagram of a GPS receiver.

three known satellite locations. The intersection of two spheres is a circle (Fig. 9.6b) and the intersection of this circle with the third sphere narrows down the location to just two points, as shown in Fig. 9.6c. One of these points is the correct location. But which one? Fortunately, one of the points gives a ridiculous answer. The incorrect point may not be on earth, or it may indicate impossibly high velocity of the receiver. The computer in a GPS receiver has various techniques for distinguishing the correct point from the incorrect one.

In practice, the timing bias α is not zero. To solve this problem, we need a distance measurement from a fourth satellite. A user locates his or her position by receiving the signal from four of the possible 24 satellites, as shown in Fig. 9.6d. There are four unknowns, the coordinates in the three-dimensional space of the user along with a timing bias within the user's receiver. These four unknowns can be solved by using four range equations to each of the four satellites.

Since DS/SS signals consist of a sequence of extremely short pulses, it is possible to measure their arrival times accurately. The GPS system can result in accuracies of 10 m anywhere on earth. The use of **differential GPS** can provide accuracy within centimeters. In this case we use one location whose position is known exactly. Comparison of its known coordinates with those read by a GPS receiver (for the same location) gives us the error (bias) of the GPS system, which can be used to correct the errors of GPS measurements of other locations. This is based on the fact that satellites are located so high that any errors measured by one receiver will be almost exactly the same for any other receiver in the same locale. Differential GPS is currently used in such diverse applications as surveying, to lay petroleum pipelines, aviation systems, marine navigation systems, preparing highly accurate maps of everything from underground electric cabling to power poles, and so on.

Why Use Spread Spectrum in GPS? The use of spread spectrum in the GPS system accomplishes three things. First, the signals from the satellites can be kept from unauthorized use. Second, and more importantly in a practical sense, the inherent processing gain of spread spectrum allows reasonable power levels to be used. Since the cost of a satellite is proportional to its weight, it is desirable to reduce the power required as much as possible. In addition, since each satellite must see the entire hemisphere, very little antenna gain is possible. For high accuracy short pulses are required to provide fine resolution. This results in high spectrum occupancy. The result is that the received signal is several decibels below the noise floor. Since range information needs to be calculated only about once every second, the data bandwidth need only be about 100 Hz. This is a natural match for spread spectrum. By despreading the received signal in the receiver, a significant processing gain is realized, thus allowing good reception at reasonable power levels. The third reason for spread spectrum is that each satellite can use the same frequency band without interfering with one another due to the near orthogonality of each user's signal.

Each satellite circles the earth in 12 hours and emits two PN sequences modulated in phase quadrature at two frequencies. Two frequencies are needed to correct for the delay introduced by the ionosphere.

9.3 TRANSMISSION MEDIA

Message-bearing electrical signals are transmitted over a distance through a variety of transmission media, ranging from a pair of wires to optical fibers, depending on the nature of the electrical signals.

An electrical signal is a form of an electromagnetic wave of a certain frequency and wavelength. Thus, a telegraph signal, a radio broadcast signal, and a light beam from the sun, a star, or a laser are all forms of electromagnetic energy of different frequencies. Various frequency bands of the electromagnetic spectrum are assigned to specific types of communication, as shown in Fig. 9.7. For each band, we need to use an appropriate transmission medium suitable for the frequency range (as indicated in Fig. 9.7). The following is a brief discussion of various transmission media encountered in practice.

9.3.1 Wire-Pair Cables

Several pairs of wires, each pair forming one transmission path, are packed in one cable. The cross talk between pairs is reduced by twisting the pairs. These are used mainly for telephone signals and low-rate data communication (see Fig. 9.7).

As the use of the telephone spread in the late nineteenth century, webs of uninsulated wire conductors supported by cross arms suspended from wooden poles began to grow through cities where telephone service was available. The conductors were insulated from the wooden support structures by glass or ceramic insulators. Circuits built by suspending such conductors on structures above ground are called **open-wire circuits**. At first, only a single such conductor was used to connect each telephone to a central switchboard, with the transmission path completed by a ground return. It quickly became evident that such ground-return circuits* were especially susceptible to interference and damage from voltages and currents induced by lightning strokes or transients in adjacent power lines. The ground return was replaced by an additional conductor, and the balanced configuration carried voice signals both to and from the telephone. The resulting circuit is called a **two-wire circuit**.†

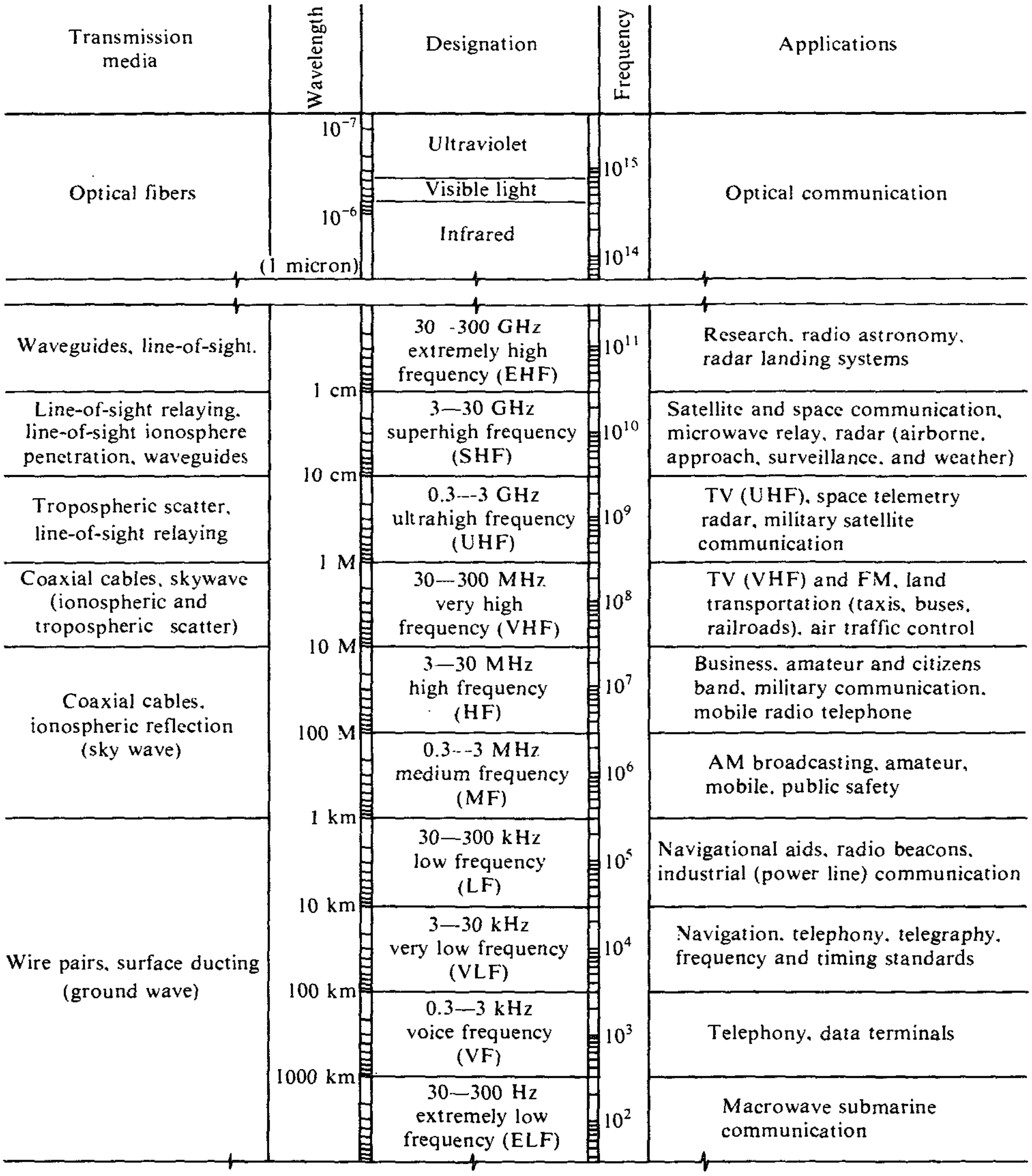
Late in the nineteenth century, methods were developed for twisting long strips of paper around conductors and then bundling such conductors inside a protective sheath. Such an assembly of conductors along with its protective sheath is called a **cable** or **telephone cable**. Pairs of wires intended to form a single circuit in such a cable are twisted slowly around each other as they pass along the cable. Typical lengths for the conductors of a pair to make a full 360 degree rotation along the axis of the cable range from about 6 to about 18 inches. The twists ensure that the conductors do not migrate away from each other and that, to the extent possible, each conductor of the pair is exposed to the same cross-talk voltages and currents from other pairs and to the same interfering voltages and currents from sources external to the cable. This minimizes induced noise and transient voltages from outside interfering sources and harmful cross-talk from other pairs within the same cable sheath.

Cable sheaths intended for use inside buildings are often formed entirely from insulating materials. Sheaths intended for use outside virtually always include a metallic covering to shield the cable pairs from external interference and to provide mechanical protection. Special steel wrappers such as "gopher tape armor" are available to prevent rodents from chewing through the usual aluminum or lead sheath material. As the art of cable building advanced, it became possible to insulate the pairs with paper pulp and solid polyethylene. During the

---

* Such circuits would be termed **common-mode** today. In modern telecommunications literature, common-mode circuits are often called **longitudinal circuits** and balanced circuits are called **metallic circuits**.

† As late as the 1960s, isolated logging and mining camps and Forest Service lookouts were served by ground-return circuits. Today most of the open-wire circuits have been replaced by metallic cables or lightwave spans. A few examples of the older technology remain in service, some carrying FDM analog carrier telephone systems.



Transmission media	Wavelength	Designation	Frequency	Applications
Optical fibers	10^{-7}	Ultraviolet	10^{15}	Optical communication
		Visible light		
	10^{-6} (1 micron)	Infrared	10^{14}	
Waveguides, line-of-sight.	1 cm	30 -300 GHz extremely high frequency (EHF)	10^{11}	Research, radio astronomy, radar landing systems
Line-of-sight relaying, line-of-sight ionosphere penetration, waveguides	10 cm	3—30 GHz superhigh frequency (SHF)	10^{10}	Satellite and space communication, microwave relay, radar (airborne, approach, surveillance, and weather)
Tropospheric scatter, line-of-sight relaying	1 M	0.3—3 GHz ultrahigh frequency (UHF)	10^{9}	TV (UHF), space telemetry radar, military satellite communication
Coaxial cables, skywave (ionospheric and tropospheric scatter)	10 M	30—300 MHz very high frequency (VHF)	10^{8}	TV (VHF) and FM, land transportation (taxis, buses, railroads), air traffic control
Coaxial cables, ionospheric reflection (sky wave)	100 M	3—30 MHz high frequency (HF)	10^{7}	Business, amateur and citizens band, military communication, mobile radio telephone
	1 km	0.3—3 MHz medium frequency (MF)	10^{6}	AM broadcasting, amateur, mobile, public safety
Wire pairs, surface ducting (ground wave)	10 km	30—300 kHz low frequency (LF)	10^{5}	Navigational aids, radio beacons, industrial (power line) communication
	100 km	3—30 kHz very low frequency (VLF)	10^{4}	Navigation, telephony, telegraphy, frequency and timing standards
	1000 km	0.3—3 kHz voice frequency (VF)	10^{3}	Telephony, data terminals
		30—300 Hz extremely low frequency (ELF)	10^{2}	Macrowave submarine communication

Figure 9.7 Various frequency bands with typical uses and transmission media.

1970s, low-loss pairs insulated with expanded polyethylene and other advanced materials were introduced.

The advent of the cable made it possible to move circuits off the elevated cross arms and place them underground,* eliminating the problems described. A large fraction of the cables placed during the early decades of this century were placed along pole lines, but such aerial cables have often been replaced by underground or buried cables in recent years.

---

* Cables suspended on poles, sometimes the same poles which formerly carried cross arms full of open-wire pairs, are generally termed **aerial cables**. Cables buried directly in the ground are often placed along rural routes by trenchers and are called **buried cables**. Cables placed under city streets in prepared ducts are called **underground cables**.

When the first digital transmission system, T1, was introduced to the Bell system in 1962, its sole application was to buried and underground cables. The power density spectrum of the alternate mark inverted (AMI) or bipolar binary line code used pulses of both polarities to represent logical **1**'s, whereas a time slot in which no pulse is transmitted represents a logical **0**. The line rate chosen was 1.544 Mbit/s. The power density of such an AMI signal peaks at about half the line rate, or 772 kHz. The first zero in the spectrum is at the frequency corresponding to the line rate, 1.544 MHz. Most of the cables for which T1 was intended had never before been used for the transmission of anything above about 300 kHz, and some special considerations were required to make the typical telephone cable useful for carrying a T1 line.

9.3.2 Coaxial Cables

At higher frequencies, wire pairs are unsuitable because they have higher electrical resistance due to the skin effect, and they suffer an increased loss of energy due to radiation from the wires. For higher frequencies, a coaxial cable is appropriate, because it eliminates both of these problems. In addition, there is virtually no cross talk between several coaxial cables that are bound together in one large cable. This is because the current in each coaxial cable is concentrated on the inside of the outer shell and the outside of the inner conductor, creating a shielding effect. For this same reason, coaxial cables are much more immune to noise and cross talk. Hence, a signal can withstand much more attenuation before requiring amplification. Whereas a single pair of wires can carry 12 or 24 voice channels, a single coaxial cable can carry 1800 to 3600 (or even more) voice channels.

9.3.3 Radio Systems

At still higher frequencies, radiation of electromagnetic waves through free space becomes attractive because of reduced antenna dimensions. The energy radiated from a transmitting antenna may reach the receiving antenna over any of several possible propagation paths, as shown in Fig. 9.8. The wave that reaches the receiving antenna after being reflected from the ionosphere is the **sky wave**. The waves that are reflected at abrupt changes in the effective dielectric constant of the troposphere (the region within 10 km of the earth's surface) are the **tropospheric waves**. Energy propagated over all other paths is considered to be the **ground wave**, which can be divided into a **space wave** and a **surface wave**. The space wave is made up of the **direct wave** (the signal that travels the direct path from transmitter to receiver) and the **ground-reflected wave**, which is the signal that arrives at the receiver after being reflected from the surface of the earth. The surface wave is a wave that is guided along the earth's surface, much as an electromagnetic wave is guided by a transmission line.

A received signal can be the resultant of a number of these waves. All these mechanisms of propagation may be present over a radio link. Some of them are negligible in certain frequency ranges, however. In the VLF band, for example, the wavelengths are so long that they are comparable to the heights of the lowest ionospheric layers (about 100 km). The ionosphere and the earth's surface act as conducting planes to form a waveguide. Thus, VLF signals can

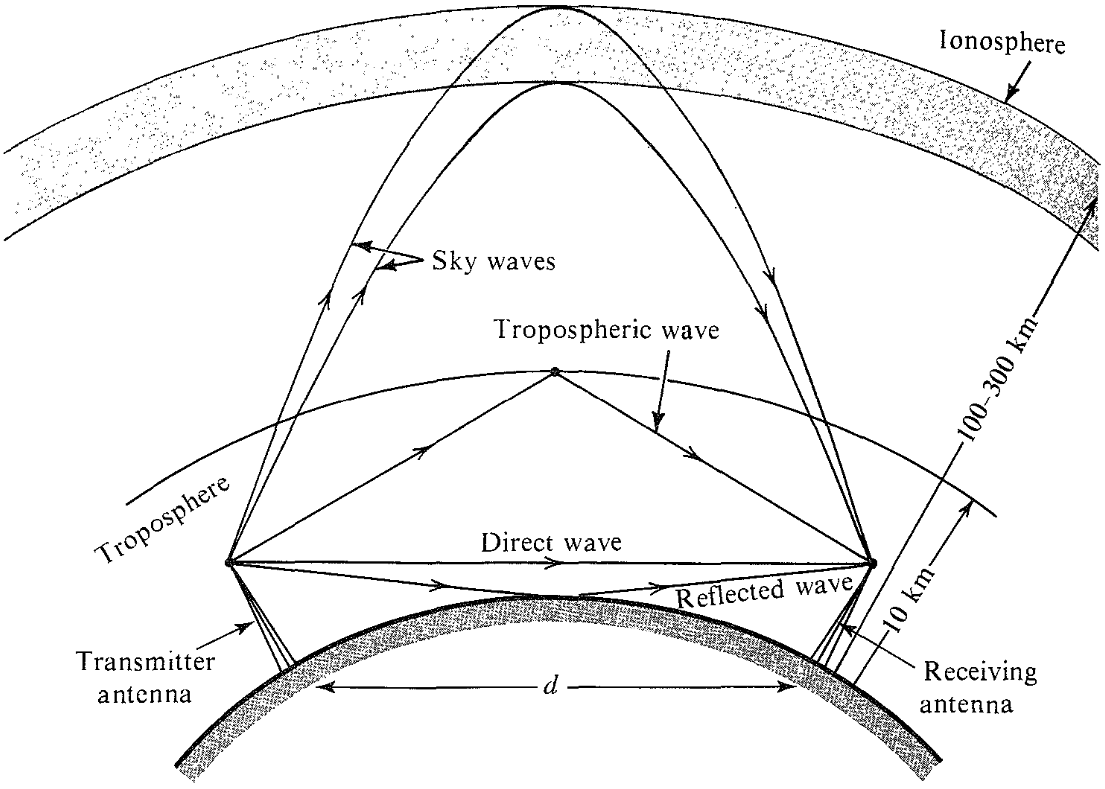


Figure 9.8 Several possible propagation paths in a radio system.

have worldwide coverage. This band is used for telegraph transmission, for navigational aids, and for distributing standard frequencies.

In the LF band, propagation is mainly from the ground wave, which provides stable transmission over distances up to about 1500 km. This band is used for long-wave sound broadcasting. In the MF and HF bands it is the sky wave that predominates. These bands are used for sound broadcasting (AM and amateur radio, etc.) and long-distance communication to ships and aircraft.

At frequencies above 30 MHz, the radio waves pass through the ionosphere instead of being reflected by it. Hence, radio communication in the VHF and UHF bands depends on the direct-wave mechanism. The range of the direct wave is the line-of-sight distance. Because of the tropospheric waves, however, the range increases beyond the optical horizon.

For line-of-sight communication the **free space loss** L, which is due to spherical dispersion of the radio waves, is given by

$$L = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi f d}{c}\right)^2$$

where d is direct distance between the transmitting and the receiving antennas, λ is the wavelength, and f is the signal frequency. The constant $c \approx 3 \times 10^8$ m/s is the velocity of light. The directional antennas, which have a focusing effect, act as amplifiers. Because of this, if P_{in} and P_{out} are the transmitted and the received powers, respectively, then

$$P_{\text{out}} = \frac{g_T g_R}{L} P_{\text{in}}$$

where g_T and g_R are the transmitter and the receiver antenna gains. The maximum gain g of a directional antenna with effective aperture area A_e is

$$g = \frac{4\pi A_e}{\lambda^2} = \frac{4\pi A_e f^2}{c^2}$$

Microwave transmission (SHF) over distances beyond the optical horizon can be obtained by means of the mechanism of **tropospheric scattering** (tropospheric reflection and refraction). Microwave transmission (along with coaxial cable) is used for bulk transmission. Of the two, microwave is the main contender. In recent years, microwave has been used more extensively than coaxial cables for the building of long-haul trunks. Like coaxial cables, microwave links today carry thousands of voice channels and are in widespread use for the transmission of television signals. They require line-of-sight transmission by a chain of relaying antennas throughout the region. Relay towers are usually spaced 30 miles apart. Thus, long-distance telephone and television signals are picked up every 30 miles, amplified, and retransmitted. A long-distance microwave circuit has fewer repeaters as compared to a coaxial cable circuit, because the repeaters are spaced about 30 miles apart in the former and 2 to 4 miles apart in the latter. This results in a superior quality signal in the microwave system as compared to the coaxial system.

During the past 20 years, extensive use has been made of the frequency bands of 2 to 10 GHz for analog signal transmission. For this reason, the development of digital microwave systems has been concentrated at higher frequencies. At present, systems operating in the 11- and 19-GHz regions are being considered. These frequencies present a severe problem of atmospheric loss of signal caused by water-vapor absorption and oxygen absorption.

9.3.4 Satellite Communication Systems

A communication satellite provides a form of microwave relay. Because it is high in the sky, its line-of-sight range is much longer, and, therefore, it relays signals over longer distances. The early communication satellites were in relatively low-altitude orbits and, consequently, sped around the earth in a few hours. This required the ground antennas to move constantly in order to beam signals to them. The satellites were overhead for only a brief period. Hence, early transatlantic television transmission was confined to a 5-min session of Walter Cronkite. In 1965, the Early Bird satellite was launched into a much higher equatorial orbit of 36,000 km (or 22,400 miles), so that it traveled around the earth once in 24 hours. Because the earth also rotates once in 24 hours, the satellite appears to be stationary over the earth. Small thrusters make adjustments to keep it as nearly stationary as possible. A growing number of stationary satellites are now being used for military and civilian purposes. Such satellites are known as **synchronous satellites**. Satellites are powered by solar batteries. The number and type of satellites to be used in a satellite relay network depends on the network coverage desired. Although three satellites in geostationary orbits (Fig. 9.9) can provide global coverage, the desire to satisfy an increased communications demand and other reasons make four or more satellites with closer spacing an obvious consideration for a global system. A smaller number of satellites is required for domestic systems that must provide regional coverage for a nation or group of nations.

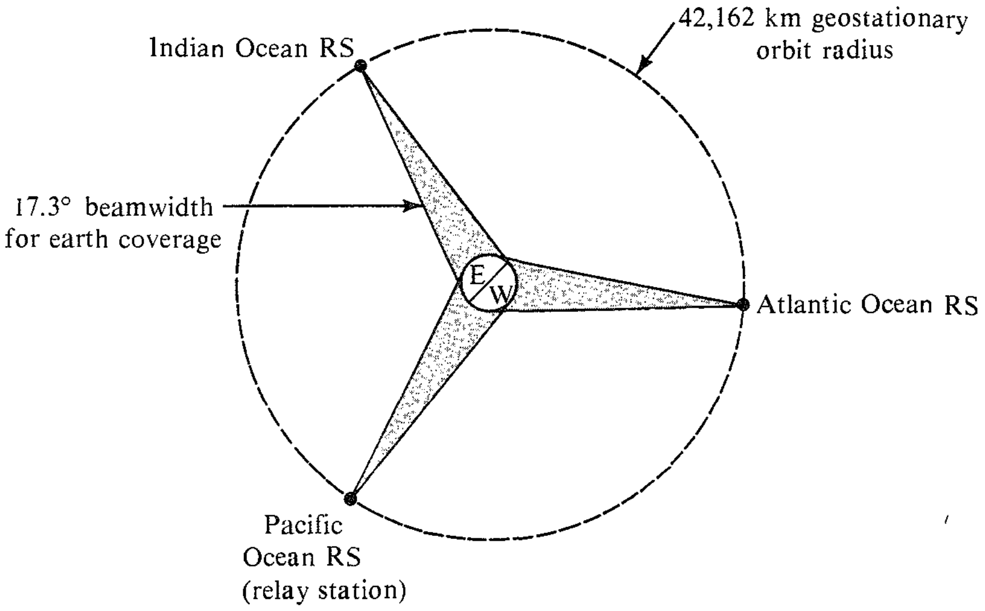

Indian Ocean RS
42,162 km geostationary orbit radius
17.3° beamwidth for earth coverage
E
W
Atlantic Ocean RS
Pacific Ocean RS (relay station)

Figure 9.9 Satellite communication.

A satellite receives the signal transmitted from the earth station, amplifies it, and then retransmits it back to earth after changing its carrier frequency. This processing is done by a transponder. Each radio channel has its own transponder. The current communication satellites typically have 24 transponders. A single transponder can carry one color television signal, 1200 voice circuits, or digital data at a rate of 50 Mbit/s.[2]

A disadvantage of a satellite link is the long propagation delay caused by great distances between the earth stations and the geostationary satellite. For a one-way channel, this delay is about 270 ms. Thus, in a two-way telephone conversation, an interval of 540 ms, more than half a second, occurs between speaking and receiving an answer. This is considerably larger than 50 ms, for the longest coast to coast terrestrial connection. Hence, the propagation delay and the associated echo become significant with satellite communications. Echo cancelers can control the echo, but nothing can be done about the delay. In practice, the delay is not objectionable to telephone users connected by a single satellite link. Connections made between two links in tandem, however, would be unsatisfactory. Consequently, the CCITT recommends that very long intercontinental connections be made over a tandem connection of a satellite link one way and a submarine cable the other way.

Satellites offer a wide bandwidth at a transmission cost that is independent of distance. As a result, they are suitable for one-way television transmission between various network locations. Almost all network and cable television broadcasters receive most of their broadcast material from satellites.

9.3.5 Optical Communication Systems

In a little over 15 years, light-wave communication using optical fibers has progressed from a laboratory proposal to a commercial reality. The widespread use of optical fibers on a routine basis began in the early eighties. Along with this has come reduced cost of existing services and the introduction of new services made more economical by this new transmission medium.

The optical band is just an extension of the radio and microwave spectrum. The laser made available a coherent optical frequency on the order of 10^{15}. Even if we use only a 0.1% bandwidth, it still corresponds to a 1000-GHz bandwidth—a transmission capacity previously undreamed of. The practical optical communication systems today operate at rates of 90 Mbit/s to 2.5 Gbit/s. But the technology is progressing rapidly. Since 1980, optical-fiber transmission capabilities have doubled every year, whereas costs have dropped exponentially. Such progress should continue beyond the year 2000. At the 1996 Optical Fiber Communications Conference in San Jose, three separate research groups (Fujitsu, Nippon Telegraph and Telephone, and AT&T) announced the successful transmission of data at a rate of 1000 Gbit/s (1 Tbit/s) over an optical fiber. This rate, considered as the holy-grail of high capacity transmission experiments, is two and a half times faster than the previous experimental record of 400 Gbit/s by Nippon. It is 400 times faster than the fastest commercial systems in use today, which carry 2.5 Gbit/s. It would take four to five years to commercialize this latest development.

This spectacular breakthrough means that we can transmit the contents of 300 years worth of daily newspapers in a single second, or convey 15 million telephone conversations (at 64 kbit/s each) simultaneously through a single optical fiber. Indeed, there might be no need to transmit so much information over a single fiber in the foreseeable future. Using bandwidths for problems such as signaling, control, and network management naturally suggests that wavelength-division multiplexing (multiplexing signals by color of light) should supersede time-division multiplexing.

Two technological developments made optical communication a reality: the development of light sources that can be modulated at high digit rates, and the production of low-loss glass fibers that act as optical waveguides. These light sources fall into two categories: **light emitting diodes (LEDs)**, which produce noncoherent light, and **lasers**, which produce coherent light. LEDs can be directly modulated by varying the drive current at a rate of up to a few hundred megabits per second. The laser is a threshold device which turns on at about 100 mA of drive current, although recent reports indicate much lower drive currents. The laser can be modulated by varying the drive current up to a rate on the order of gigabits per second.

Two general classes of receivers are used to recover the optical signal from the fiber and convert it back into an electrical signal. The PIN diode includes an intrinsic layer between a p-doped layer and an n-doped layer. Photons striking the intrinsic layer create hole-electron pairs which allow current flow in the external circuit connected to the p and n layers. The **avalanche photodiode (APD)** works in a similar manner, except that the bias across the junction is high enough to accelerate holes and electrons to velocities that will produce more hole-electron pairs in collisions. This multiplicative feature adds gain to the detection process. For longer wavelengths, germanium and indium gallium arsenide detectors are being considered. These detectors convert light into electrical currents with excellent efficiency and low noise.

The optical fibers produced today have low losses (a fraction of a decibel per kilometer), exceptionally low bit-error rates on the order of 10^{-11} or lower, an absence of cross talk, fewer repeaters, very large capacity, and intrinsic compatibility with digital transmission. All this enables a repeater spacing of 40 km to be commonplace today, and this spacing is increasing with the use of improved high-power InGaAsP lasers and higher sensitivity receivers employing PIN-FET and avalanche photodetectors.

Telecommunication on optical fiber has grown rapidly in recent years. Several transcontinental single-mode fiber links have already been completed in the United States and Canada. In the United States, fiber transmission systems are being installed by long-distance carriers such as AT&T, Sprint, MCI, Lightnet, the National Telecommunication Network Group (NTN), and regional Bell companies and by independently operating telephone companies.

Several submarine fiber-optic cables (TAT-12/13 between the United States, Britain, and France and TPC-5 CN between California, Hawaii, Japan, and Guam) are also completed. The longest human-made structure ever being assembled is the world's most ambitious undersea light-wave communication system between Europe and Japan via the Mediterranean and the Indian Ocean, linking the United Kingdom, Spain, Italy, Egypt, United Arab Emirates, India, Malaysia, Thailand, Hongkong, China, Korea (north and south), and Japan. When completed in 1998, the system will be using complex undersea optical-fiber cable that will span 27,300 km –more than two thirds of the earth's circumference. Called the **fiber-optic link around the globe (FLAG)**, it will snake in eight sections through the Atlantic Ocean, the Mediterranean, the Red Sea, the Indian Ocean, and the Pacific Ocean. The first length was installed at the end of 1995, and early 1998 will see completion.

The backbone of the FLAG is third-generation transoceanic optical-fiber cable technology. The first two generations carried up to 280 and 560 Mbit/s of data per pair of optical fibers, respectively; FLAG raises the rate to 5.3 Gbit/s. The system will be able to carry 120,000 circuits as 64-kbit/s channels on two fiber pairs. In comparison, in 1956 the first transatlantic telephone copper cable carried only 36 conversations, whereas the first optical-fiber cable installed in 1988 across the Atlantic Ocean carried 8000 circuits, also as 64-kbit/s channels on two fiber pairs.

The dramatic advances in fiber optics are making satellite communication obsolete, at least for large-bandwidth point-to-point communication systems such as transoceanic telephone systems. Just 30 years ago satellite systems were thought to be the communication systems of the future.

The optical-fiber systems have increased the communication capacity tremendously, and as the long-haul market becomes saturated, telephone companies are planning to replace twisted copper wire circuits with optical fiber between local central offices and subscribers' homes. This is already being done in some residential communities in Europe and Japan.

A Historical Note *

Optical transmission has a long and interesting history. Alexander Graham Bell built a primitive system in which a vibrating mirror modulated a light beam. During the late 1800s, the U.S. Army used an optical telegraph system powered by sunlight to communicate across the vast distances of the Southwest. During World War II, the U.S. Navy modulated light beams using vacuum tube technology. Such systems were useful for providing secure ship-to-ship communication.

During the 1960s, short-range optical systems of low bandwidth were built in which the optical signal was guided by a rather thick, high-loss optical fiber. The fiber was usually formed of plastic. Such systems were used for shipboard communication, to reduce the weight of aircraft by eliminating heavy copper cables, to illuminate automobile dashboards, and for a host of other applications where the lengths of fiber were so short that the high attenuation and the large pulse dispersion of the fiber were not important. Such systems still see wide use today.

---

* Much of the remaining chapter was contributed by Mr. Maynard Wright.

Many of these early systems used (and still use) lightwaves in the visible region of the spectrum. For telecommunications, the infrared spectrum is virtually always used. For the remainder of this section, and in the telecommunications literature in general, the term optical should be taken as referring to signals in the infrared spectrum.

The predominance of infrared technology in telecommunications results from the fact that certain bands of frequencies in the infrared range are subject to much lower attenuation in relatively pure glass than are visible signals or signals in other portions of the infrared region. Most optical transmission research and production have centered on these low-loss bands.

During the late 1960s and early 1970s, advances were made in the production of glass fibers which made it possible to form a very thin fiber of high mechanical strength and, more important from a transmission standpoint, great purity. The high purity produces a very low attenuation per unit length since most of the signal lost in transmission along the fiber is lost by absorption in impurities in the glass.

Early fibers were made with a very thin core of circular cross section surrounded by a thicker layer of differing refractive index. The outer layer in such a "step index" fiber is termed the **cladding**. The refractive indices of the core and the cladding are chosen so that lightwaves transported through the core which are not parallel with the central axis of the core are reflected away from the cladding and back into the core. When lightwaves are launched into such a fiber by the transmitting device, the waves assume any of a continuum of angles whose span is governed by the focusing mechanism of the transmitter. Those that are beyond the critical angle for the materials of the core and the cladding pass into the cladding and are lost. Those that are under the critical angle are reflected back and forth, from the junction with the cladding first on one side of the core and then on the other. Waves that happen to be launched parallel to the core will, in the case of a perfectly straight fiber, pass directly to the distant receiver with no reflections at all.

The reflected waves pass from the transmitter to the receiver via a longer path than do the direct waves, the increase in distance being a function of the angle with which a particular wave is launched into the end of the fiber (Fig. 9.10a). The tiny diameter of the core compared to the length of the fiber means that the percent difference between maximum and minimum distances traveled by the various component waves is very small. The difference may, nevertheless, be significant over a long distance to cause significant dispersion of a pulse waveform, especially if the signal has a very high pulse repetition rate. This form of distortion of the signal is termed **multimode dispersion.**[*] Variation with frequency of the refractive index of the fiber produces a form of dispersion called **material dispersion**. Variation with frequency of the ratio of the refractive indices of the core and cladding leads to **profile dispersion**, and **waveguide dispersion** results from geometric effects on phase and group velocities as a function of frequency.

A more advanced fiber known as the **graded index** fiber has a refractive index which varies with the distance from the core in a manner that causes waves traveling along paths which diverge from the centerline of the fiber to be "herded" back toward the center, as shown in Fig. 9.10b. A wave propagated down such a core therefore oscillates about the centerline in a sinusoidal manner rather than by traveling the reflected straight-line segments of the step index core. Such fibers also have velocities of propagation that are higher at greater distances from the center of the core. The arrival times of waves traveling over paths of different lengths are thus equalized to a certain extent by the graded index fiber. Such fibers were used to build fiber routes ranging from 45 up to 135 Mbit/s or more during the 1970s and 1980s.

---

[*] Multimode dispersion is more properly termed "multimode distortion" since the effect is due to the combination at the receiver of waves having traveled via various paths, each of which may have been dispersion-free. The effect is therefore not directly dependent on the dispersive qualities of the glass.

Most new fiber placement today utilizes single-mode fibers. A single-mode fiber has a core with such a small diameter that only one waveguide mode may be propagated through it (Fig. 9.10c). This eliminates multimode dispersion entirely. Such a fiber is still subject to the effects of the frequency-dependent dispersion mechanisms.

A typical optical transmitter used with multimode fibers, such as an LED, will produce a signal containing a broad range of optical frequencies. To take advantage of the capabilities of the single-mode fiber, a narrow-band laser transmitter is required. The use of such a narrow-band transmitter to produce a very narrow range of frequencies at the input to the fiber reduces significantly the other forms of dispersion and allows relatively high signaling rates of many gigabits per second to be used.

The length of fiber between the transmitter and the receiver is limited by the attenuation of the fiber and/or a combination of the various dispersions discussed. An additional limitation is loss and reflection in splices. Unlike metallic cable pairs, in which splices generally (but not always) contribute negligibly to the loss and phase distortion of the pairs, splices in optical fibers may be responsible for the majority of the distortion between the transmitter and the receiver of a typical link.

Two general forms of splice are used: a mechanical connector and a fusion splice. In the mechanical connector, the two fiber ends to be joined are cut square and polished and then held in close alignment by a special connector assembly. Index matching fluid may be introduced to fill the very small air gap between portions of the two machined surfaces.

In the fusion splice, the two ends to be joined are held in close alignment by a tool and then melted together using an electric arc to form a "welded" joint. Either type of splice may produce more attenuation than many kilometers of optical fiber since each introduces

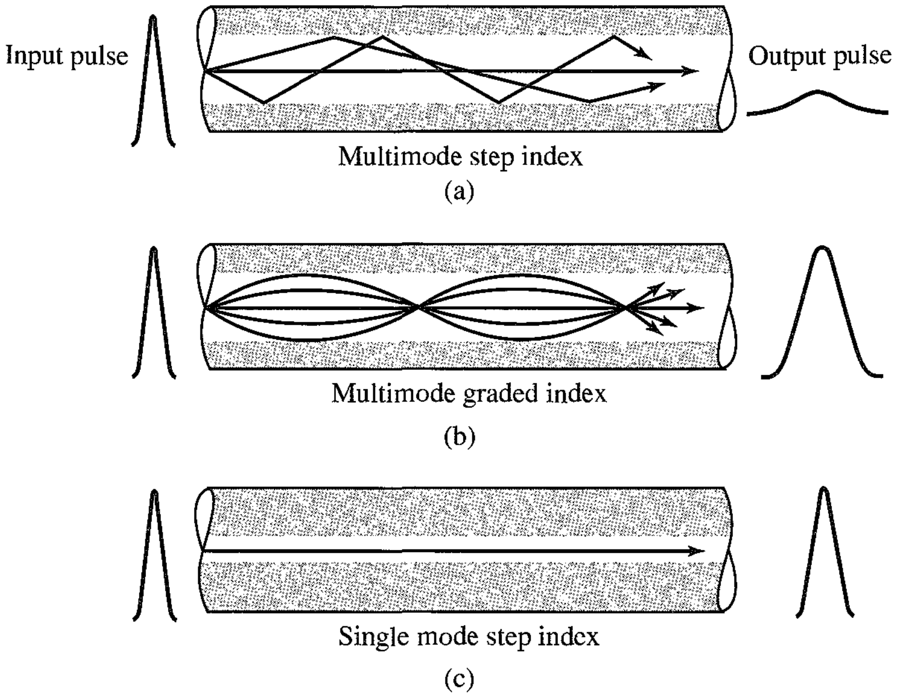


Figure 9.10 Signal transmission through optical fiber. (a) Multimode fiber. (b) Graded fiber. (c) Single-mode fiber.

a substantial impedance discontinuity. It is not possible to predict with accuracy the effect of an individual splice. The splices must therefore be included in transmission calculations in a statistical manner. This differs from the deterministic methodology used throughout transmission calculations for metallic cable.

Note that, in optical transmission, it is the power of the optical waveform that is modulated by the signal to be transported and not the amplitude of a current or voltage. This is a necessary constraint imposed by the transmitter and the receiver and by the nature of the optical signal, which occupies a band of frequencies rather than a single sinusoidal carrier, as is common in radio transmission. Modulation of the power always results in a positive signal level, and there is no direct optical equivalent of the bipolar electrical signal. It is possible, however, to map the various levels of a bipolar or multilevel signal into an equivalent number of optical power levels. The advantages of binary logic and its ease of implementation in optical hardware have made binary transmission virtually universal in optical telecommunication systems.

Although most modern optical communication links utilize optical fibers, commercial systems are available which implement point-to-point optical transmission paths through the atmosphere. These links suffer more drastically than do comparable microwave radio installations from degradation due to fog, flocks of birds, or other disturbances of the transmission path. The optical systems, however, can be installed rapidly, require no FCC licensing of the transmitter, and are largely immune to interference from (and to) other installations. This makes them superior to microwave systems for certain short-term applications such as broadcasts from sporting events or emergency restoration of routes failed by natural disasters.

9.4 HYBRID CIRCUIT: 2-WIRE TO 4-WIRE CONVERSIONS

Almost all telephone instruments are connected to the serving central office by a single cable pair which carries both directions of transmission. The amplifiers processing such signals, however, will generally pass signals in only one direction. The same is true of the telephone instrument itself, which includes a transmitter and a receiver along with separate circuitry for each. Hence, some technique for converting the mode of transmission from combined to separate must be provided. The circuit in which both directions of transmission flow in a single pair is termed a **2-wire** circuit and that in which the directions of transmission (transmit and receive) are separated and flow in two different pairs is known as a **4-wire** circuit (Fig. 9.11).

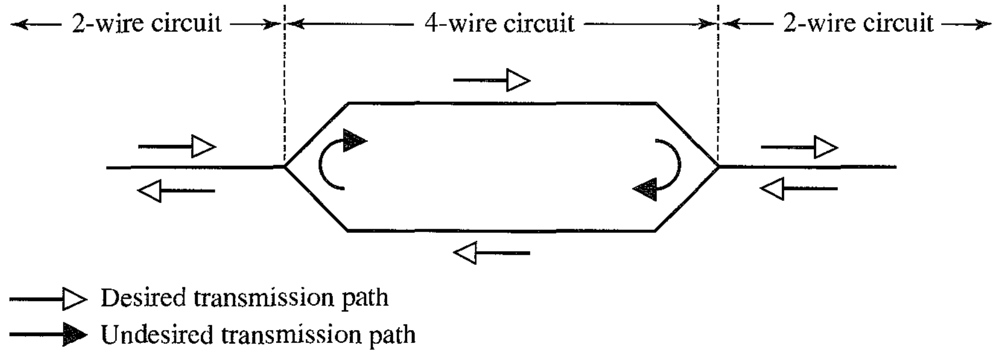


Figure 9.11 Two-wire and four-wire transmission paths.

At first glance, it might seem to be adequate to simply connect the two paths in parallel at any junction between 2-wire and 4-wire paths. Such a connection is shown in Fig. 9.11. If one were careful to match impedances, there are points where this technique would work, although not well for several reasons. In most cases, however, the 4-wire path exhibits positive gain in both directions. To tie the inputs and outputs together at each end would produce a closed loop with gain (positive feedback) at at least some frequencies. Such a structure would oscillate, obliterating the signals to be passed and possibly creating harmful cross talk interference to other circuits.

A circuit is needed, therefore, which will pass energy from the 2-wire circuit into the transmit side of the 4-wire circuit and energy from the receive side of the 4-wire circuit into the 2-wire circuit without allowing energy to flow from the receive side of the 4-wire circuit into the transmit side of the 4-wire circuit. Such a circuit is called a **4-wire terminating** set. There are two general classes of such circuits: resistive hybrids and hybrid coil sets.

A resistive hybrid consists of a Wheatstone bridge (Fig. 9.12a) with four equal resistive legs, which is connected to the 4-wire circuit in such a way that the current induced in the circuit by the receive side is balanced out in the transmit side and no energy flows from one 4-wire path to the other. Each of the equal resistances is set equal to the impedance of the 2-wire path to which the circuit is to be connected. One of the resistances is then replaced by the 2-wire circuit (Fig. 9.12b). Since the 2-wire circuit has the same impedance as the resistance it replaces, the circuit is still balanced with respect to the two 4-wire ports. The 2-wire port is balanced with respect to neither of the 4-wire ports and energy can flow between the 2-wire port and either of the 4-wire ports. There will be some loss in the process due to the resistance of the network, but it is made up by amplifiers in the circuit.

The circuit will not be perfectly balanced because the 2-wire circuit will generally exhibit an impedance which is a function of frequency and which may vary somewhat from the ideal or design value. As long as the losses around the loop formed by the two halves of the 4-wire circuit exceed by a reasonable margin the gains in the same portions of the circuit, the circuit will perform properly. When a loud squeal is heard during a telephone conversation, the problem may often be traced to some impedance mismatch which allows excessive energy to flow across the undesired path through a hybrid. When a circuit is not

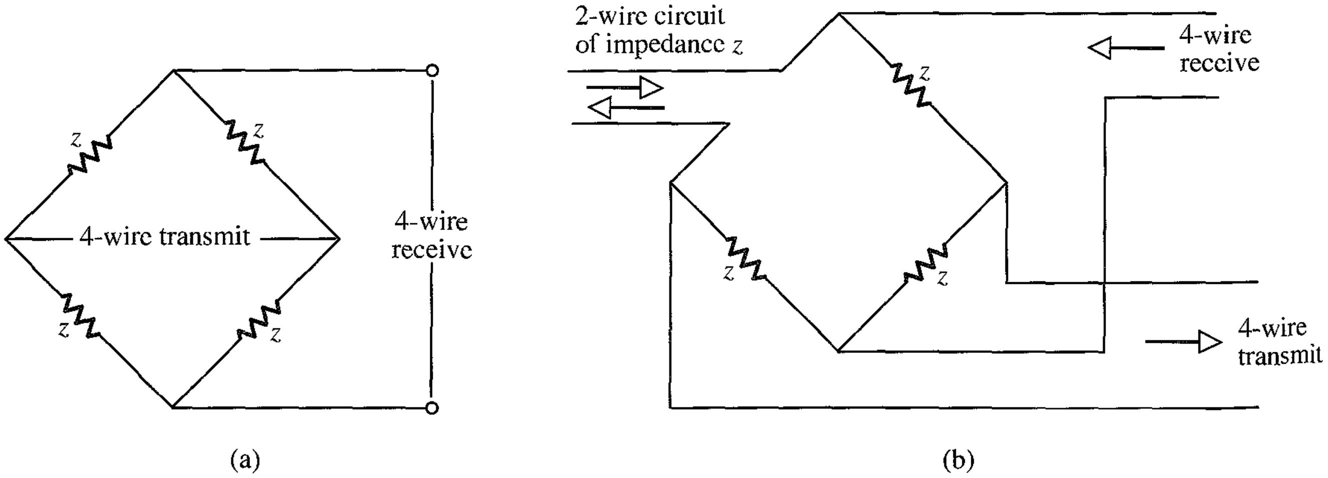


Figure 9.12 Wheatstone bridge hybrid circuit.

quite in the oscillatory state but approaches it, it may sound very hollow and produce excessive echoing of voice signals.

Various methods for combining transformers into assemblies called hybrid coils are also used to achieve the 2-wire to 4-wire conversion. The high quality transformers required make hybrid coils more expensive than resistive hybrids. Hybrid coils, however, introduce less loss into a circuit than do resistive hybrids. There are several possible ways to arrange transformers to form hybrid coil sets. One common arrangement is shown in Fig. 9.13. The hybrid coil set shown is formed from four independent but identical transformers, usually iron-core units for voice frequency applications. They are labeled T_1 through T_4 in Fig. 9.13 for the purposes of the following explanation, but they are usually manufactured and installed as a single unit. As in the case of the resistive hybrid, a perfect balance requires the impedance z to be equal to the impedance of the cable pair, and it will be assumed here that such is the case. For clarity, the terms primary and secondary will not be used, but windings will be identified by reference to their transformers and by use of the adjectives "upper" and "lower" in reference to Fig. 9.13.

When current flows in the upper windings of T_1 and T_2 due to the arrival of energy from the receiving 4-wire port, the voltages induced in the lower windings of the two transformers will cause current to flow in the 2-wire circuit and in impedance z. Since both have the same impedance, the same current will flow. Energy is thus transferred from the 4-wire circuit into the 2-wire circuit. Note that at least half the energy received from the 4-wire circuit is lost in impedance z since, by symmetry, the same amount of energy is transferred into z and into the 2-wire circuit. This 3-dB loss is an unavoidable by-product of the use of the hybrid coil sets.

The currents flowing in the 2-wire circuit and in z will also flow through the upper windings of T_3 and T_4. By symmetry of T_1 and T_2 and due to the fact that the same current flows through both their upper windings, the currents flowing through the upper windings of T_3 and T_4 will be equal, and the voltages induced in their lower windings will therefore be equal and in the same direction in the figure. Note that the lower windings of T_3 and T_4 are connected in opposition so that the two voltages will cancel and produce no current flow in the 4-wire transmit circuit. Energy transfer between the 4-wire receive port and the 4-wire transmit port is thus inhibited.

When energy flows into the hybrid coil set from the 2-wire circuit, current is caused to flow in the lower winding of T_1 and in the upper winding of T_3. The induced voltages cause currents to flow into both 4-wire ports. The reversed winding of T_4 causes the voltages appearing across impedance z to be zero and no energy is expended in z. The energy flowing into

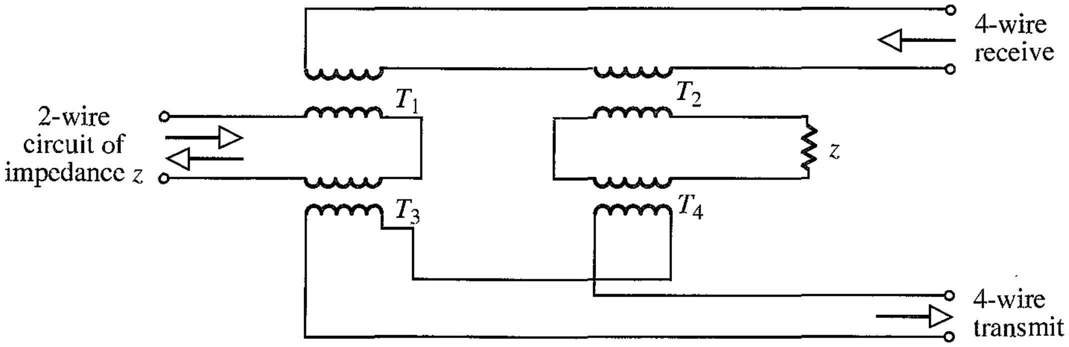


Figure 9.13 Balanced coil set hybrid circuit.

the receive 4-wire port is undesired, but it is flowing into the output circuitry of a transmission facility designed to carry energy in the opposite direction, and it will normally be harmlessly dissipated in the output impedance of an amplifier. As in the case of energy transfer in the opposite direction, the energy directed to the desired port is reduced by 3 dB by the division of energy.

The transformers associated with the hybrid coil sets are not lossless and will add at least a few tenths of a decibel of loss. The 3-dB loss in each direction is therefore a lower theoretical limit.

4-wire terminating sets of both types have been used for decades to allow voice frequency circuits to employ one cable pair for both directions of transmission. Recent advances in echo canceler technology have made it possible to combine both directions of transmission of a digital system into a single pair, and hybrids are also required for that application. Systems using such technology include the high bit rate digital subscriber lines (HDSL) and certain of the integrated services digital network (ISDN) local loop architectures.

Time-Assignment System Interpolation

Long-distance speech communication uses separate paths (two one-way circuits) for the two directions (send and receive) of transmission. Because, on the average, a talker talks half the time and listens half the time, each of the circuits will be used half the time only. Measurements on working transatlantic lines show that speech activity is present only about 40% of the time on each circuit. This means each circuit is free about 60% of the time, on the average. This fact is used to interpolate additional talkers onto communication facilities using a **time-assignment system interpolation (TASI)** concept.[3] TASI systems have mainly been used with submarine-cable voice-channel facilities. New systems may operate with mixed satellite and cable voice channels. TASI is a voice-operated switching and channel-assignment system. A new incoming speech signal is assigned to a channel that is temporarily unused by another talker. The new talker will keep the channel until he or she is silent and the channel is needed for another talker. During the low-traffic period, the system may require no switching, and the talker may keep the same channel throughout the conversation. During high-traffic periods, however, successive portions of the same talker's speech signal may be assigned and switched to different channels. Assignment and switching are accomplished in milliseconds, so that little of the initial syllable is lost.

Using the TASI system, a large number of customers can be served by a smaller number of channels. For example, the TASI B system uses 96 channels to serve 235 customers.

9.5 PUBLIC SWITCHED TELEPHONE NETWORK (PSTN)

Initial applications of the telephone in the late nineteenth century involved permanently wired connections between two telephone instruments.* Sometimes signaling arrangements were included to "ring" the phone at one end when the other party wished to communicate. When more than two parties wished to be able to communicate, the station sets involved were connected in parallel via the wires that ran between the station locations.

It quickly became inconvenient to simply tie every station together in parallel as the number of stations increased. The initial solution, which persisted into the middle years of this century and has not entirely disappeared yet, was to establish a switchboard at which manual connections of various stations could be made. Arrangements were made to signal the operator of the switchboard when someone wished to place a call. The calling party would ask the operator of the switchboard to connect his or her line to the called party's line. When one or the other of the parties to the call completed the call by hanging up, the operator would see an indication and would remove the cords that established the call.

---

* A telephone instrument is often called a **station set** and that terminology will be used here.

Even though this made it possible to give each station a separate line to the switchboard, it was economically unattractive to do so in rural areas where running the wire was expensive and difficult, especially in the days prior to cable, when each separate station required an individual circuit involving two wires and numerous insulators on cross arms. A single circuit was often, therefore, run along a rural road and numerous stations were "bridged" (connected in parallel) on the circuit. Such circuits were known as party lines, as opposed to private lines which carried only a single station.

Late in the nineteenth century, methods were developed in this country and in Europe for allowing customers to switch their own calls. Dials* were added to station sets as switching machines, which could respond to the digits dialed from the stations, replaced operator boards. This change increased the privacy of communication by taking the telephone operator out of the circuit† and made it possible to handle the enormous increase in telephone usage which occurred as the telephone went from novelty to useful tool to necessity in the years bracketing the turn of the century.

Until after the middle of this century, local calls could be dialed from the station, but long-distance calls continued to be handled by the operator. Switching systems were relatively simple, and the network consisted of stations, the local loop plant,‡ local switching machines, operator switchboards, and the interoffice transmission systems and equipment.

The initial switching systems were **step-by-step (SXS)** machines. They responded to dialed numbers by setting or "stepping" multiple-contact magnet-driven switches to connect two stations through a sequence of connections involving a matrix of such switches. They were relatively slow and expensive, but they were installed up into the 1960s, and some remain in service today.§

During the 1950s, switching machines called **common control switchers** were introduced by Western Electric and others. These machines included various complex elements which would carry out the several functions involved in establishing a call, leave the call set up through a matrix of simple switches called **crossbar** switches, and then become available for use in routing another call. These machines used a much less expensive switching matrix than did the step-by-step machines, and the complex equipment required for call setup was made reusable prior to the termination of each call, introducing economies of scale that were unavailable in earlier machines.

The introduction of common control switching and improvements in transmission technology made possible the introduction of **direct distance dialing (DDD)** during the 1950s, further reducing the role of the telephone operator in telecommunications. The advent of DDD made it necessary to organize the network of switching machines employed by the Bell system into a logical hierarchy. Switching machines were divided into five different classes. The lowest class (class 5) or "end office" machines connected stations together. Each switch "homed" on another switch one level higher in the hierarchy. A station placing a call to another station would be switched through progressively higher class switching machines and would eventually reach a point in the hierarchy at which both stations homed on the same switch. Connections would then be made down through the hierarchy to the called station. The exception here would be when the two stations homed on different class 1 (regional) switches. In that case, the call would be routed between the two involved regional switches in the same manner as between the other switches in the connection.

---

* The signaling device on a pretone signaling telephone is almost always called a **dial**. Its proper name is **finger wheel**.

† Almon B. Strowger, a Kansas City undertaker, is credited with developing the first telephone automatic switching system in 1891 to prevent telephone operators from allegedly diverting business to his competitors.[4]

‡ The **loop** or **local loop** is the open-wire pair or cable pair that connects the central office to the subscriber's station.

§ Switching machines are frequently called **switchers** or just **switches**. All three designations will be used in this text as the context best fits.

The circuits connecting two switching machines provided a transmission path, a signaling path, and circuitry to connect the transmission path to each switch. These elements taken together form an interswitch facility called a **trunk** or a **trunk circuit**.

When any of various factors, such as the existence of a commercial community of interest, caused a relatively high volume of calls to be placed between two switches not connected by a direct group of trunks, such a group might be built. The trunk group would consist of high usage trunks if it provided the most direct path between two particular switching machines or of "final" trunks if the group was a part of the hierarchy between the five classes of switches.

The breakup of the Bell system in the early 1980s changed the nature of the telephone network considerably. The seven regional Bell operating companies (BOCs) were separated from AT&T. Each region was carved up into sections called "exchange areas" or "local access and transport areas" (LATAs). BOCs could switch and transport telephone traffic and data within LATAs but could not carry traffic across LATA boundaries, even if the two or more LATAs involved were in the same BOC's territory. Conversely, AT&T and the other interexchange carriers could not transport traffic within LATAs. This court-mandated change is often referred to as "divestiture."

The massive reorganization associated with divestiture sparked many changes in the ways that telephone traffic, including voice and data services, is both switched and transported. Local calls and close regional calls will generally be handled by switches and transmission equipment belonging to a BOC or one of the many independent local exchange companies (LECs) that were unaffiliated with the Bell system even prior to divestiture. If the call (or nonswitched voice or data service) must cross a LATA boundary, the BOC or LEC at one end must hand the signal off to an interexchange carrier (IEC) for transport across the boundary. The IEC will then hand the signal back to a BOC or LEC (which may or may not be the same company that originated the signal), which will deliver it to the end user or switch.

These changes have given rise to an entirely new standards organization, the Exchange Carrier Standards Association (ECSA). ECSA has recently changed its name to the Alliance for Telecommunications Industry Solutions (ATIS). This organization consists of representatives from all segments of the industry and creates standards that are approved by the American National Standards Institute (ANSI). The formal American national standards produced by ATIS supersede the internal specifications and requirements of the old Bell system and are intended to establish the ground rules for signals and protocols at the interfaces between the various types of service providers involved in today's competitive telecommunications environment.

REFERENCES

1. Committee on Digital CDMA Cellular, *IS-95 Wideband Spread Digital Cellular System Dual Mode Mobile Station-Base Station Compatibility Standard*, Technical Report, EIA/TIA, TR 45.5, April 1992.
2. John G. Nellist, *Understanding Telecommunications and Lightwave Systems*, IEEE Press, New York, 1992.
3. J. M. Fraser, D. B. Bullock, and N. G. Long, "Overall Characteristics of a TASI System," *Bell Syst. Tech. J.*, vol. 51, pp. 1439–1454, July 1962.
4. John Brooks, *Telephone*, Harper & Row, New York, 1975.

10 INTRODUCTION TO THEORY OF PROBABILITY

Thus far, we have studied signals whose values at any instant t were known from their analytical or graphical description. Such signals are called **deterministic** signals, implying complete certainty about their values at any instant t. Such signals, which can be specified with certainty, cannot convey information. It will be seen in Chapter 15 that information is related to uncertainty. The higher the uncertainty about a signal (or message) to be received, the higher its information content. If a message to be received is specified (i.e., if it is known beforehand), it has no uncertainty and, consequently, cannot convey any information. Hence, signals that convey information must be unpredictable. Noise signals that perturb information signals are also unpredictable. These unpredictable message signals and noise waveforms are examples of **random processes**.

Random phenomena arise either because of our partial ignorance of the generating mechanism (as in message or noise signals) or because the laws governing the phenomena may be fundamentally random (as in quantum mechanics). Yet in another situation, such as the outcome of rolling a die, it is possible to predict the outcome provided we know exactly all the conditions, such as the angle of the throw, the nature of the surface on which it is thrown, the force imparted by the player, and so on. The exact analysis, however, is so complex that it is impractical to carry it out, and we are content to accept the outcome prediction on an average basis. Here the random phenomenon arises from our unwillingness to carry out exact analysis because it is not worth the trouble.

We shall begin with a review of the basic concepts of the theory of probability, which forms the basis for describing random processes.

10.1 CONCEPT OF PROBABILITY

We begin by defining some important terms. An experiment is called a **random experiment** if its outcome cannot be predicted precisely because the conditions under which it is performed cannot be predetermined with sufficient accuracy and completeness. Tossing a coin, rolling a die, and drawing a card from a deck are some examples of random experiments. A random experiment may have several separately identifiable **outcomes**. For example, rolling a die has six possible identifiable outcomes (1, 2, 3, 4, 5, and 6). **Events** are sets of outcomes meeting some specifications. In the experiment of rolling a die, for example, the event "odd number on a throw" can result from any one of three outcomes (viz., 1, 3, and 5). Hence, this event is a set of three outcomes (1, 3, and 5). Thus, events are groupings of outcomes into classes among which we choose to distinguish. These ideas can be better understood by using the concepts of set theory.

We define the **sample space** S as a collection of all possible separately identifiable outcomes of a random experiment. Each outcome is an **element**, or **sample point**, of this space and can be conveniently represented by a point in the sample space. In the random experiment of rolling a die, for example, the sample space consists of six elements represented by six sample points ζ_1, ζ_2, ζ_3, ζ_4, ζ_5, and ζ_6, where ζ_i represents the outcome "a number i is thrown" (Fig. 10.1). The event, on the other hand, is a subset of S. The event "an odd number is thrown," denoted by A_o, is a subset of S (or a set of sample points ζ_1, ζ_3, and ζ_5). Similarly, the event A_e, "an even number is thrown," is another subset of S (or a set of sample points ζ_2, ζ_4, and ζ_6),

$$A_o = (\zeta_1,\ \zeta_3,\ \zeta_5) \qquad\qquad A_e = (\zeta_2,\ \zeta_4,\ \zeta_6)$$

Let us denote the event "a number equal to or less than 4 is thrown" as B. Thus, $B = (\zeta_1,\ \zeta_2,\ \zeta_3,\ \zeta_4)$. These events are clearly marked in Fig. 10.1. Note that an outcome can also be an event, because an outcome is a subset of S with only one element.

The **complement** of any event A, denoted by A^c, is the event containing all points not in A. Thus, for the event B in Fig. 10.1, $B^c = (\zeta_5,\ \zeta_6)$, $A_o^c = A_e$, and $A_e^c = A_o$. An event that has no sample points is a **null event**, which is denoted by $\emptyset$ and is equal to S^c.

The **union** of events A and B, denoted by $A \cup B$, is that event which contains all points in A and B. This is the event "A or B." For the events in Fig. 10.1,

$$A_o \cup B = (\zeta_1,\ \zeta_3,\ \zeta_5,\ \zeta_2,\ \zeta_4)$$
$$A_e \cup B = (\zeta_2,\ \zeta_4,\ \zeta_6,\ \zeta_1,\ \zeta_3)$$

Observe that

$$A \cup B = B \cup A \tag{10.1}$$

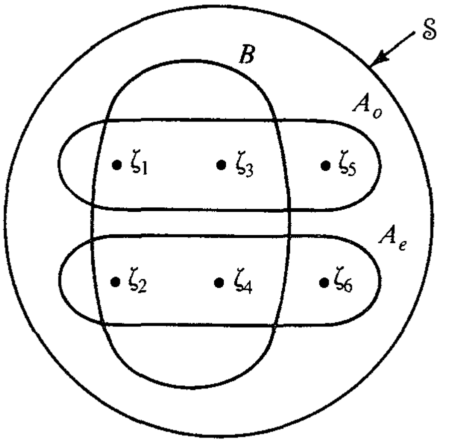


Figure 10.1 Sample space for a throw of a die.

The **intersection** of events A and B, denoted by $A \cap B$ or simply AB, is the event that contains points common to A and B. This is the event "both A and B," also known as the **joint event** AB. Thus, the event A_eB, "a number that is even and equal to or less than 4 is thrown," is a set $(\zeta_2,\ \zeta_4)$, and similarly for A_oB,

$$A_eB = (\zeta_2,\ \zeta_4), \qquad A_oB = (\zeta_1,\ \zeta_3)$$

Observe that

$$AB = BA \tag{10.2}$$

All these concepts can be demonstrated on a Venn diagram (Fig. 10.2). If the events A and B are such that

$$AB = \emptyset \tag{10.3}$$

then A and B are said to be **disjoint**, or **mutually exclusive**, events. This means events A and B cannot occur simultaneously. In Fig. 10.1 events A_e and A_o are mutually exclusive, meaning that in any trial of the experiment if A_e occurs, A_o cannot occur at the same time, and vice versa.

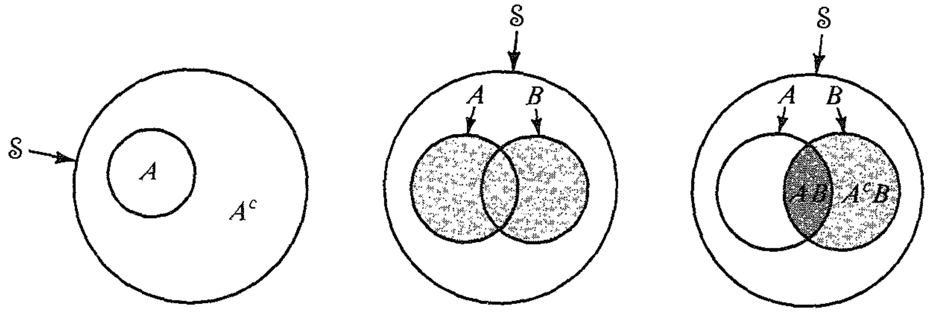
Relative Frequency and Probability

Although the outcome of a random experiment is unpredictable, there is a *statistical regularity* about the outcomes. For example, if a coin is tossed a large number of times, about half the times the outcome will be "heads," and the remaining half of the times it will be "tails." We may say that the relative frequency of the two outcomes "heads" or "tails" is one-half.

Let A be one of the events of a random experiment. If we conduct a sequence of N independent trials* of this experiment, and if the event A occurs in $N(A)$ out of these N trials, then the fraction

$$f(A) = \lim_{N\to\infty} \frac{N(A)}{N} \tag{10.4}$$

is called the **relative frequency** of the event A. Observe that for small N, the fraction $N(A)/N$ may vary widely with N. As N increases, the fraction will approach a limit because of statistical regularity.



A^c
(a)

$A \cup B = B \cup A$
(b)

AB and A^cB
(c)

Figure 10.2 Representation of complement, union, and intersection of events.

---

[*] Trials conducted under similar discernible conditions.

The probability of an event has the same connotations as the relative frequency of that event. Hence, to each event we assign a probability that is equal to the relative frequency of that event.* Therefore, to an event A, we assign the probability $P(A)$ as

$$P(A) = \lim_{N\to\infty} \frac{N(A)}{N} \tag{10.5}$$

From Eq. (10.5), it follows that

$$0 \le P(A) \le 1 \tag{10.6}$$

EXAMPLE 10.1 Assign probabilities to each of the six outcomes in Fig. 10.1.

Because each of the six outcomes is equally likely in a large number of independent trials, each outcome will appear in one-sixth of the trials. Hence,

$$P(\zeta_i) = \frac{1}{6} \qquad i = 1,\ 2,\ 3,\ 4,\ 5,\ 6 \tag{10.7}$$

Consider now the two events A and B of a random experiment. Suppose we conduct N independent trials of this experiment and events A and B occur in $N(A)$ and $N(B)$ trials, respectively. If A and B are mutually exclusive (or disjoint), then if A occurs, B cannot occur, and vice versa. Hence, the event $A \cup B$ occurs in $N(A) + N(B)$ trials and

$$\begin{aligned} P(A \cup B) &= \lim_{N\to\infty} \frac{N(A) + N(B)}{N} \\ &= P(A) + P(B) \qquad \text{if } AB = \emptyset \end{aligned} \tag{10.8}$$

This result can be extended to more than two mutually exclusive events.

EXAMPLE 10.2 Assign probabilities to the events A_e, A_o, B, A_eB, and A_oB in Fig. 10.1.

Because $A_e = (\zeta_2 \cup \zeta_4 \cup \zeta_6)$ where ζ_2, ζ_4, and ζ_6 are mutually exclusive,

$$P(A_e) = P(\zeta_2) + P(\zeta_4) + P\zeta_6)$$

From Eq. (10.7) it follows that

$$P(A_e) = \frac{1}{2} \tag{10.9a}$$

Similarly,

$$P(A_o) = \frac{1}{2} \tag{10.9b}$$

$$P(B) = \frac{2}{3} \tag{10.9c}$$

From Fig. 10.1 we also observe that

$$A_eB = \zeta_2 \cup \zeta_4$$

---

* Observe that we are not *defining* the probability by the relative frequency. To a given event, we *assign* a probability that is equal to the relative frequency of the event. Modern theory of probability, being a branch of mathematics, starts with certain axioms about probability [Eqs. (10.6), (10.8), and (10.11)]. It does not concern itself with how the probability is assigned to an event. It assumes that somehow these probabilities are assigned. We use relative frequency to assign probability because it is reasonable in the sense that it closely approximates our expectation of "probability."

and

$$P(A_eB) = P(\zeta_2) + P(\zeta_4) = \frac{1}{3} \tag{10.10a}$$

Similarly,

$$P(A_oB) = \frac{1}{3} \tag{10.10b}$$

We can also show that

$$P(S) = 1 \tag{10.11}$$

This result can be proved using the relative frequency. Let a random experiment be repeated N times (N large). Because S is the union of all possible outcomes, S occurs in every trial. Hence, N out of N trials are favorable to S and the result follows.

EXAMPLE 10.3 Two dice are thrown. Determine the probability that the sum on the dice is seven.

For this experiment, the sample space contains 36 sample points because 36 possible outcomes exist. All the outcomes are equally likely. Hence, the probability of each outcome is 1/36.

A sum of seven can be obtained by the six combinations: (1, 6), (2, 5), (3, 4), (4, 3) (5, 2), and (6, 1). Hence, the event "a seven is thrown" is the union of six outcomes, each with probability 1/36. Therefore,

$$P(\text{“a seven is thrown”}) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}$$

EXAMPLE 10.4 A coin is tossed four times in succession. Determine the probability of obtaining exactly two heads.

A total of $2^4 = 16$ distinct outcomes are possible, all of which are equally likely because of the symmetry of the situation. Hence, the sample space consists of 16 points, each with probability 1/16. The 16 outcomes are as follows:

HHHH	TTTT
HHHT	TTTH
HHTH	TTHT
⟶HHTT	⟶TTHH
HTHH	THTT
⟶HTHT	⟶THTH
⟶HTTH	⟶THHT
HTTT	THHH

Six out of these 16 outcomes are favorable to the event "obtaining exactly two heads" (shown by arrows). Because all of the six outcomes are disjoint (mutually exclusive),

$$P(\text{obtaining exactly two heads}) = \frac{6}{16} = \frac{3}{8}$$

In example 10.4, the method of listing all possible outcomes quickly becomes unwieldy as the number of tosses increases. For example, if a coin is tossed just 10 times, the total number of outcomes is 1024. A more convenient approach would be to apply the results of combinatorial analysis used in Bernoulli trials, to be discussed shortly.

Conditional Probability and Independent Events

Conditional Probability: One often comes across a situation where the probability of one event is influenced by the outcome of another event. As an example, consider drawing two cards in succession from a deck. Let A denote the event that the first card drawn is an ace. We do not replace the card drawn in the first trial. Let B denote the event that the second card drawn is an ace. It is evident that the probability of drawing an ace in the second trial will be influenced by the outcome of the first draw. If the first draw does not result in an ace, then the probability of obtaining an ace in the second trial is 4/51. The probability of event B thus depends on whether or not event A occurs. We now introduce the **conditional probability** $P(B|A)$ to denote the probability of event B when it is known that event A has occurred. $P(B|A)$ is read as "probability of B given A."

Let an experiment be performed N times, in which the event A occurs n_1 times. Of *these* n_1 trials, event B occurs n_2 times. It is clear that n_2 is the number of times that the joint event AB (Fig. 10.2c) occurs. That is,

$$P(AB) = \lim_{N\to\infty}\left(\frac{n_2}{N}\right) = \lim_{N\to\infty}\left(\frac{n_1}{N}\right)\left(\frac{n_2}{n_1}\right)$$

Note that $\lim_{N\to\infty}(n_1/N) = P(A)$. Also, $\lim_{N\to\infty}(n_2/n_1) = P(B|A)$,* because B occurs n_2 of the n_1 times that A occurred. This represents the conditional probability of B given A. Therefore,

$$P(AB) = P(A)P(B|A) \tag{10.12}$$

and

$$P(B|A) = \frac{P(AB)}{P(A)} \qquad \text{provided } P(A) \neq 0 \tag{10.13a}$$

Using a similar argument, we obtain

$$P(A|B) = \frac{P(AB)}{P(B)} \qquad \text{provided } P(B) \neq 0 \tag{10.13b}$$

It follows from Eqs. (10.13) that

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{10.14a}$$

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} \tag{10.14b}$$

Equations (10.14) are called **Bayes' rule**. In Bayes' rule, one conditional probability is expressed in terms of the reversed conditional probability.

---

* Here we are implicitly using the fact that $n_1 \to \infty$ as $N \to \infty$. This is true provided the ratio $\lim_{N\to\infty}(n_1/N) \neq 0$ [that is, if $P(A) \neq 0$].

EXAMPLE 10.5 A random experiment consists of drawing two cards from a deck in succession (without replacing the first card drawn). Assign a value to the probability of obtaining two red aces in two draws.

Let A and B be the events "red ace in the first draw" and "red ace in the second draw," respectively. We wish to determine $P(AB)$,

$$P(AB) = P(A)P(B|A)$$

and the relative frequency of A is $2/52 = 1/26$. Hence,

$$P(A) = \frac{1}{26}$$

Also, $P(B|A)$ is the probability of drawing a red ace in the second draw given that the first draw was a red ace. The relative frequency of this event is $1/51$, so

$$P(B|A) = \frac{1}{51}$$

Hence,

$$P(AB) = \left(\frac{1}{26}\right)\left(\frac{1}{51}\right) = \frac{1}{1326}$$

Independent Events: Under conditional probability, we presented an example where the occurrence of one event was influenced by the occurrence of another. There are, of course, many examples where two or more events are entirely independent; that is, the occurrence of one event in no way influences the occurrence of the other event. As an example, we again consider the drawing of two cards in succession, but in this case we replace the card obtained in the first draw and shuffle the deck before the second draw. In this case, the outcome of the second draw is in no way influenced by the outcome of the first draw. Thus $P(B)$, the probability of drawing an ace in the second draw, is independent of whether or not the event A (drawing an ace in the first trial) occurs. Thus, the events A and B are independent. The conditional probability $P(B|A)$ is given by $P(B)$.

The event B is said to be **independent** of the event A if

$$P(B|A) = P(B) \tag{10.15a}$$

It can been seen from Eqs. (10.14) that if event B is independent of event A, then event A is also independent of B; that is,

$$P(A|B) = P(A) \tag{10.15b}$$

Note that if the events A and B are independent, it follows from Eqs. (10.13a) and (10.15a) that

$$P(AB) = P(A)P(B) \tag{10.15c}$$

Bernoulli Trials

In Bernoulli trials, if a certain event A occurs, we call it a "success." If $P(A) = p$, then the probability of success is p. If q is the probability of failure, then $q = 1 - p$. We shall find the probability of k successes in n (Bernoulli) trials. The outcome of each trial is independent of the outcomes of the other trials. It is clear that in n trials, if success occurs in k trials, failure occurs in $n-k$ trials. Since the outcomes of the trials are independent, the probability of this event is clearly $p^n(1-p)^{n-k}$, that is,

$$P(k \text{ successes in a specific order in } n \text{ trials}) = p^k(1-p)^{n-k}$$

But the event of "k successes in n trials" can occur in many different ways (different orders). It is well known from the combinatorial analysis that there are $\binom{n}{k}$ ways in which k things can be taken from n things (which is the same as the number of ways of achieving k successes in n trials), where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{10.16}$$

This can be proved as follows. Consider an urn containing n distinguishable balls marked $1, 2, \ldots, n$. Suppose we draw k balls from this urn without replacement. The first ball could be any one of the n balls, the second ball could be any one of the remaining $(n-1)$ balls, and so on. Hence, the total number of ways in which k balls can be drawn is

$$n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!}$$

Next, consider any one set of the k balls drawn. These balls can be ordered in different ways. We could choose any one of the k balls for number 1, and any one of the remaining $(k-1)$ balls for number 2, and so on. This will give a total of $k(k-1)(k-2)\cdots 1 = k!$ distinguishable patterns formed from the k balls. The total number of ways in which k things can be taken from n things is $n!/(n-k)!$ But many of these ways will use the same k things, arranged in different order. The ways in which k things can be taken from n things without regard to order (unordered subset k taken from n things) is $n!/(n-k)!$ divided by $k!$ This is precisely $\binom{n}{k}$ defined by Eq. (10.16).

This means the probability of k successes in n trials is

$$\begin{aligned} P(k \text{ successes in } n \text{ trials}) &= \binom{n}{k} p^k(1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \end{aligned} \tag{10.17}$$

Tossing a coin and observing the number of heads is a Bernoulli trial with $p = 0.5$. Hence, the probability of observing k heads in n tosses is

$$P(k \text{ heads in } n \text{ tosses}) = \binom{n}{k}(0.5)^k(0.5)^{n-k} = \frac{n!}{k!(n-k)!}(0.5)^n$$

EXAMPLE 10.6 A binary symmetric channel (BSC) has an error probability P_e (i.e., the probability of receiving **0** when **1** is transmitted, or vice versa, is P_e). Note that the channel behavior is symmetrical with respect to **0** and **1**. Thus,

$$P(\mathbf{0}|\mathbf{1}) = P(\mathbf{1}|\mathbf{0}) = P_e$$

and

$$P(\mathbf{0}|\mathbf{0}) = P(\mathbf{1}|\mathbf{1}) = 1 - P_e$$

where $P(y|x)$ denotes the probability of receiving y when x is transmitted. A sequence of n binary digits is transmitted over this channel. Determine the probability of receiving exactly k digits in error.

The reception of each digit is independent of the other digits. This is an example of a Bernoulli trial with the probability of success $p = P_e$ ("success" here is receiving a digit in error). Clearly, the probability of k successes in n trials (k errors in n digits) is

$$P(\text{receiving } k \text{ out of } n \text{ digits in error}) = \binom{n}{k} P_e^k(1 - P_e)^{n-k}$$

For example, if $P_e = 10^{-5}$, the probability of receiving two digits wrong in a sequence of eight digits is

$$\binom{8}{2}(10^{-5})^2(1 - 10^{-5})^6 \simeq \frac{8!}{2!\,6!}10^{-10} = (2.8)10^{-9}$$

EXAMPLE 10.7 PCM Repeater Error Probability

In PCM, regenerative repeaters are used to detect pulses (before they are lost in noise) and retransmit new, clean pulses. This combats the accumulation of noise and pulse distortion.

A certain PCM channel consists of n identical links in tandem (Fig. 10.3). The pulses are detected at the end of each link and clean new pulses are transmitted over the next link. If P_e is the probability of error in detecting a pulse over any one link, show that P_E, the probability of error in detecting a pulse over the entire channel (over the n links in tandem), is

$$P_E \simeq nP_e \qquad nP_e \ll 1$$

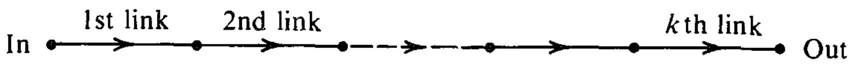


Figure 10.3 A PCM repeater.

The probabilities of detecting a pulse correctly over one link and over the entire channel (n links in tandem) are $1 - P_e$ and $1 - P_E$, respectively. A pulse can be detected correctly over the entire channel if either the pulse is detected correctly over every link or errors are made over an even number of links only.

$$\begin{aligned}1 - P_E &= P(\text{correct detection over all links})\\ &\quad + P(\text{error over two links only})\\ &\quad + P(\text{error over four links only}) + \cdots\\ &\quad + P(\text{error over } \alpha \text{ links only})\end{aligned}$$

where α is n or $n - 1$, depending on whether n is even or odd. Because pulse detection over each link is independent of the other links (see Example 10.6),

$$P(\text{correct detection over all } n \text{ links}) = (1 - P_e)^n$$

and

$$P(\text{error over } k \text{ links only}) = \frac{n!}{k!(n-k)!}P_e^k(1 - P_e)^{n-k}$$

Hence,

$$1 - P_E = (1 - P_e)^n + \sum_{k=2,4,6,\ldots}^{\alpha} \frac{n!}{k!(n-k)!} P_e^k (1 - P_e)^{n-k}$$

In practice, $P_e \ll 1$, so only the first two terms on the righthand side of this equation are of significance. Also, $(1 - P_e)^{n-k} \simeq 1$, and

$$\begin{aligned} 1 - P_E &\simeq (1 - P_e)^n + \frac{n!}{2!(n-2)!} P_e^2 \\ &= (1 - P_e)^n + \frac{n(n-1)}{2} P_e^2 \end{aligned}$$

If $nP_e \ll 1$, then the second term can also be neglected, and

$$\begin{aligned} 1 - P_E &\simeq (1 - P_e)^n \\ &\simeq 1 - nP_e \qquad nP_e \ll 1 \end{aligned}$$

and

$$P_E \simeq nP_e$$

We can explain this result heuristically by considering the transmission of N ($N \to \infty$) pulses. Each link makes NP_e errors, and the total number of errors is approximately nNP_e (approximately, because some of the errored pulses over a link will be errored over other links). Thus the overall error probability is nP_e.

EXAMPLE 10.8 In binary communication, one of the techniques used to increase the reliability of a channel is to repeat a message several times. For example, we can send each message (**0** or **1**) three times. Hence, the transmitted digits are **000** (for message **0**) or **111** (for message **1**). Because of channel noise, we may receive any one of the eight possible combinations of three binary digits. The decision as to which message is transmitted is made by the majority rule; that is, if at least two of the three detected digits are **0**, the decision is **0**, and so on. This scheme permits correct reception of data even if one out of three digits is in error. Detection error occurs only if at least two out of three digits are received in error. If P_e is the error probability of one digit, and $P(\epsilon)$ is the probability of making a wrong decision in this scheme, then

$$\begin{aligned} P(\epsilon) &= \sum_{k=2}^{3} \binom{3}{k} P_e^k (1 - P_e)^{3-k} \\ &= 3P_e^2(1 - P_e) + P_e^3 \end{aligned}$$

In practice, $P_e \ll 1$, and

$$P(\epsilon) \simeq 3P_e^2$$

For instance, if $P_e = 10^{-4}$, $P(\epsilon) \simeq 3 \times 10^{-8}$. Thus, the error probability is reduced from 10^{-4} to 3×10^{-8}. We can use any odd number of repetitions for this scheme to function.

In this example, higher reliability is achieved at the cost of a reduction in the rate of information transmission by a factor of 3. We shall see in Chapter 15 that more efficient ways exist to trade off between reliability and the rate of transmission.

Axiomatic Theory of Probability

The relative frequency definition of probability has great intuitive appeal. Unfortunately, it has some serious mathematical objections. Logically there is no reason why we should get the same estimate of the relative frequency whether we base it on 10,000 trials or on 20. Moreover, in the relative frequency definition, it is not clear when and in what mathematical sense the limit in Eq. (10.5) exists. If we consider a set of an infinite number of trials, we can partition such a set into several subsets, such as odd and even numbered trials. Each of these subsets (of infinite trials each) would have its own relative frequency. So far, all the attempts to prove that the relative frequencies of all the subsets are equal have proved futile.[1] There are some other difficulties also. For instance, in some cases, such as Julius Caesar having visited Great Britain, we cannot repeat the experiment an infinite number of times, so we can never know the probability of such an event. We, therefore, need to develop a theory of probability which is not tied down to any particular definition of probability. In other words, we must separate the empirical and the formal problems of probability. Assigning probabilities to events is an empirical aspect, and setting up purely formal calculus to deal with probabilities (assigned by whatever empirical method) is the formal aspect.

It is instructive to consider here the basic difference between physical sciences and mathematics. Physical sciences are based on **inductive logic**, and mathematics is strictly a **deductive logic**. Inductive logic consists of making a large number of observations and then generalizing from these observations, laws that will explain these observations. For instance, millions of years of experience and observation tells us that every human being must die someday. This leads to a law that *humans are mortals*. This is inductive logic. Based on a law (or laws) obtained by inductive logic, we can make further deductions, such as, John is a human being, so he must die some day. This is deductive logic. Derivation of laws of physical sciences is basically an exercise in inductive logic, whereas mathematics is pure deductive logic. In a physical science we make observations in a certain field, and generalize these observations in laws such as Ohm's law, Maxwell's equations, quantum mechanics, and so on. There are no proofs for these inductively obtained laws. They are found to be true by observation. But once we have such inductively formulated laws (axioms or hypotheses), using thought process, we can deduce additional results based on these laws or axioms alone. This is the proper domain of mathematics. All these deduced results have to be proved rigorously based on a set of axioms. Thus, based on Maxwell's equations alone, we can derive the laws of propagation of electromagnetic waves.

This discussion shows that the discipline of mathematics can be summed up in one aphorism "This implies that". In other words, if we are given a certain set of axioms (hypotheses), then, based upon these axioms alone, what else is true? As Bertrand Russell puts it: "Pure mathematics consists entirely of asseverations as that, if such and such proposition is true of anything, then such and such another proposition is true of that thing." Seen in this light, it may appear that assigning probability to an event may not necessarily be the responsibility of the mathematical discipline of probability. As a mathematical discipline, we need to start with a set of axioms about probability, and then investigate what else can be said about probability based on this set of axioms alone. We start with a concept (as yet undefined) of probability and postulate axioms. The axioms must be internally consistent and should conform to the observed relationships and behavior of probability in the practical and the intuitive sense. It is beyond the scope of this book to discuss how these axioms are formulated. The modern theory of probability starts with Eqs. (10.6), (10.8), and (10.11) as its axioms. Based on these three axioms alone, what else is true is the essence of modern theory of probability. The relative frequency approach uses Eq. (10.5) to define probability, and Eqs. (10.5), (10.8), and (10.11) follow as a consequence of this definition. In the axiomatic approach, on the other hand, we do not say anything about how we assign probability $P(A)$ to an event A, but we postulate that the probability function must obey the three postulates or axioms in Eqs. (10.6), (10.8), and (10.11). The modern theory of probability does not concern itself with the problem of assigning probabilities to events. It assumes that somehow the probabilities are assigned to these events on an a priori basis.

If a mathematical model is to conform to the real phenomenon, we must assign these probabilities consistent with an empirical and an intuitive understanding of probability. The relative frequency is admirably suited for this. Thus, although we use relative frequency to assign (not define) probabilities, it is all under the table, not a part of the mathematical discipline of probability.

10.2 RANDOM VARIABLES

The outcome of a random experiment may be a real number (as in the case of rolling a die), or it may be nonnumerical and describable by a phrase (such as "heads" or "tails" in tossing a coin). From a mathematical point of view, it is desirable to have numerical values for all outcomes. For this reason, we assign a real number to each sample point according to some rule. If there are m sample points $\zeta_1, \zeta_2, \ldots, \zeta_m$, then using some convenient rule, we assign a real number $x(\zeta_i)$ to sample point ζ_i $(i = 1, 2, \ldots, m)$. In the case of tossing a coin, for example, we may assign the number 1 for the outcome heads and the number -1 for the outcome tails (Fig. 10.4).

Thus, x(.) is a function that maps sample points $\zeta_1, \zeta_2, \ldots, \zeta_m$ into real numbers $x_1, x_2, \ldots, x_n$.* We now have a **random variable** x that takes on values $x_1, x_2, \ldots, x_n$. We shall use Roman type (x) to denote a random variable (RV) and italic type (for example, $x_1, x_2, \ldots, x_n$, etc.) to denote the value it takes. The probability of an RV x taking a value x_i is $P_{\mathrm{x}}(x_i)$.

Discrete Random Variables

A random variable is discrete if there exists a denumerable sequence of distinct numbers x_i such that

$$\sum_i P_{\mathrm{x}}(x_i) = 1 \tag{10.18}$$

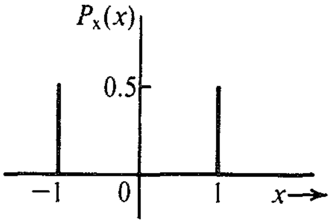


Figure 10.4 Probabilities in a coin tossing experiment.

---

* m is not necessarily equal to n. More than one sample point can map into one value of x.

Thus, a discrete RV can assume only certain discrete values. An RV that can assume any value from a continuous interval is called a **continuous** random variable.

EXAMPLE 10.9 Two dice are thrown. The sum of the points appearing on the two dice is an RV x. Find the values taken by x, and the corresponding probabilities.

x can take on all integral values from 2 through 12. Various probabilities can be determined by the method outlined in Example 10.3.

There are 36 sample points in all, each with probability 1/36. Dice outcomes for various values of x are shown in Table 10.1. Note that although there are 36 sample points, they all map into 11 values of x. This is because more than one sample point maps into the same value of x. For example, six sample points map into $x = 7$.

The reader can verify that $\sum_{i=2}^{12} P_{\mathrm{x}}(x_i) = 1$.

Table 10.1

Value of x x_i	**Dice Outcomes**	$P_{\mathrm{x}}(x_i)$
2	(1, 1)	1/36
3	(1, 2), (2, 1)	2/36 = 1/18
4	(1, 3), (2, 2), (3, 1)	3/36 = 1/12
5	(1, 4), (2, 3), (3, 2), (4, 1)	4/36 = 1/9
6	(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)	5/36
7	(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)	6/36 =1/6
8	(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)	5/36
9	(3, 6), (4, 5), (5, 4), (6, 3)	4/36 = 1/9
10	(4, 6), (5, 5), (6, 4)	3/36 = 1/12
11	(5, 6), (6, 5)	2/36 = 1/18
12	(6, 6)	1/36

The preceding discussion can be extended to two RVs x and y. The joint probability $P_{\mathrm{xy}}(x_i,\ y_j)$ is the probability that $\mathrm{x} = x_i$ and $\mathrm{y} = y_j$. Consider, for example, the case of a coin tossed twice in succession. If the outcomes of the first and second tosses are mapped into RVs x and y, then x and y each take values 1 and -1. Because the outcomes of the two tosses are independent, x and y are independent, and

$$P_{\mathrm{xy}}(x_i,\ y_j) = P_{\mathrm{x}}(x_i)\ P_{\mathrm{y}}(y_j)$$

and

$$P_{\mathrm{xy}}(1,\ 1) = P_{\mathrm{xy}}(1,\ -1) = P_{\mathrm{xy}}(-1,\ 1) = P_{\mathrm{xy}}(-1,\ -1) = \tfrac{1}{4}$$

These probabilities are plotted in Fig. 10.5.

For a general case where the variable x can take values $x_1,\ x_2,\ \ldots,\ x_n$ and the variable y can take values $y_1,\ y_2,\ \ldots,\ y_m$, we have

$$\sum_i \sum_j P_{\mathrm{xy}}(x_i,\ y_j) = 1 \qquad (10.19)$$

This follows from the fact that the summation on the left is the probability of the union of all possible outcomes and must be unity (a certain event).

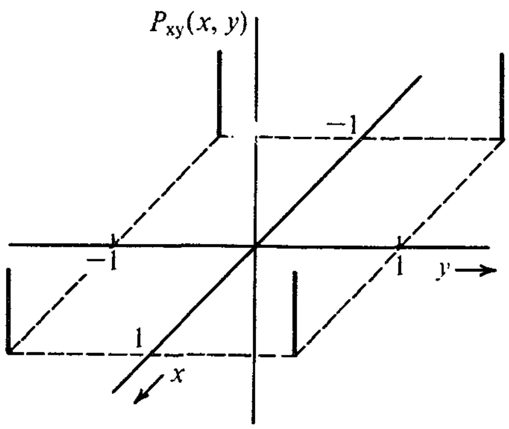


Figure 10.5 Representation of joint probabilities of 2 random variables.

Conditional Probabilities

If x and y are two RVs, then the conditional probability of $x = x_i$ given $y = y_j$ is denoted by $P_{x|y}(x_i|y_j)$. Moreover,

$$\sum_i P_{x|y}(x_i|y_j) = \sum_j P_{y|x}(y_j|x_i) = 1 \tag{10.20}$$

This can be proved by observing that probabilities $P_{x|y}(x_i|y_j)$ are specified over the sample space corresponding to the condition $y = y_j$. Hence, $\sum_i P_{x|y}(x_i|y_j)$ is the probability of the union of all possible outcomes of x (under the condition $y = y_j$) and must be unity (a certain event). A similar argument applies to $\sum_j P_{y|x}(y_j|x_i)$. Also from Eq. (10.12), we have

$$P_{xy}(x_i, y_j) = P_{x|y}(x_i|y_j)P_y(y_j) = P_{y|x}(y_j|x_i)P_x(x_i) \tag{10.21}$$

Bayes' rule follows from Eq. (10.21). Also from Eq. (10.21), we have

$$\begin{aligned}\sum_i P_{xy}(x_i, y_j) &= \sum_i P_{x|y}(x_1|y_j)P_y(y_j)\\ &= P_y(y_j)\sum_i P_{x|y}(x_i|y_j)\\ &= P_y(y_j)\end{aligned} \tag{10.22a}$$

Similarly,

$$P_x(x_i) = \sum_j P_{xy}(x_i, y_j) \tag{10.22b}$$

The probabilities $P_x(x_i)$ and $P_y(y_j)$ are called **marginal probabilities**. Equations (10.22) show how to determine marginal probabilities from joint probabilities. Results of Eqs. (10.19) through (10.22) can be extended to more than two RVs.

EXAMPLE 10.10 A binary-symmetric channel (BSC) error probability is P_e. The probability of transmitting **1** is Q, and that of transmitting **0** is $1 - Q$ (Fig. 10.6). Determine the probabilities of receiving **1** and **0** at the receiver.

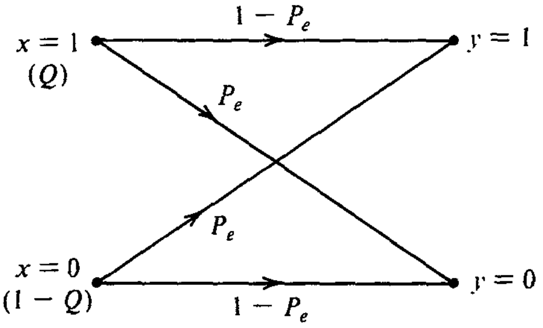


Figure 10.6 Binary-symmetric channel (BSC).

If x and y are the transmitted digit and the received digit, respectively, then for a BSC,

$$P_{y|x}(\mathbf{0}|\mathbf{1}) = P_{y|x}(\mathbf{1}|\mathbf{0}) = P_e$$

$$P_{y|x}(\mathbf{0}|\mathbf{0}) = P_{y|x}(\mathbf{1}|\mathbf{1}) = 1 - P_e$$

Also,

$$P_x(\mathbf{1}) = Q \qquad \text{and} \qquad P_x(\mathbf{0}) = 1 - Q$$

We need to find $P_y(\mathbf{1})$ and $P_y(\mathbf{0})$. Because

$$\begin{aligned} P_y(y_j) &= \sum_i P_{xy}(x_i,\ y_j) \\ &= \sum_i P_x(x_i)P_{y|x}(y_j|x_i) \end{aligned}$$

$$\begin{aligned} P_y(\mathbf{1}) &= P_x(\mathbf{0})P_{y|x}(\mathbf{1}|\mathbf{0}) + P_x(\mathbf{1})P_{y|x}(\mathbf{1}|\mathbf{1}) \\ &= (1 - Q)P_e + Q(1 - P_e) \end{aligned}$$

Similarly, we find

$$P_y(\mathbf{0}) = (1 - Q)(1 - P_e) + QP_e$$

These answers seem almost obvious from Fig. 10.6.

Note that because of channel errors, the probability of receiving a digit **1** is not the same as that of transmitting **1**. The same is true of **0**.

EXAMPLE 10.11 Over a certain binary communication channel, the symbol **0** is transmitted with probability 0.4 and **1** is transmitted with probability 0.6. It is given that $P(\epsilon|\mathbf{0}) = 10^{-6}$ and $P(\epsilon|\mathbf{1}) = 10^{-4}$, where $P(\epsilon|x_i)$ is the probability of detecting the error given that x_i is transmitted. Determine $P(\epsilon)$, the error probability of the channel.

If $P(\epsilon,\ x_i)$ is the joint probability that x_i is transmitted and it is detected wrongly, then [Eq. (10.22b)]

$$\begin{aligned} P(\epsilon) &= \sum_i P(\epsilon,\ x_i) \\ &= P(\epsilon,\ \mathbf{0}) + P(\epsilon,\ \mathbf{1}) \end{aligned}$$

$$
\begin{aligned}
&= P_{\mathrm{x}}(\mathbf{0})P(\epsilon|\mathbf{0}) + P_{\mathrm{x}}(\mathbf{1})P(\epsilon|\mathbf{1}) \\
&= 0.4(10^{-6}) + 0.6(10^{-4}) \\
&= 0.604(10^{-4})
\end{aligned}
$$

Note that $P(\epsilon|0) = 10^{-6}$ means that on the average, one out of 1 million received **0**'s will be detected erroneously. Similarly, $P(\epsilon|1) = 10^{-4}$ means that on the average, one out of 10,000 received **1**'s will be in error. But $P(\epsilon) = 0.604(10^{-4})$ indicates that on the average, one out of $1/0.604(10^{-4}) \simeq 16{,}556$ digits (regardless of whether they are **1**'s or **0**'s) will be received in error.

Cumulative Distribution Function

The **cumulative distribution function (CDF)** $F_{\mathrm{x}}(x)$ of an RV x is the probability that x takes a value less than or equal to x; that is,

$$F_{\mathrm{x}}(x) = P(\mathrm{x} \leq x) \tag{10.23}$$

We can show that a CDF $F_{\mathrm{x}}(x)$ has the following four properties:

1. $F_{\mathrm{x}}(x) \geq 0$ (10.24a)
2. $F_{\mathrm{x}}(\infty) = 1$ (10.24b)
3. $F_{\mathrm{x}}(-\infty) = 0$ (10.24c)
4. $F_{\mathrm{x}}(x)$ is a nondecreasing function, that is,

$$F_{\mathrm{x}}(x_1) \leq F_{\mathrm{x}}(x_2) \text{ for } x_1 \leq x_2 \tag{10.24d}$$

The first property is obvious. The second and third properties are proved by observing that $F_{\mathrm{x}}(\infty) = P(\mathrm{x} \leq \infty)$ and $F_{\mathrm{x}}(-\infty) = P(\mathrm{x} \leq -\infty)$. To prove the fourth property, we have, from Eq. (10.23),

$$
\begin{aligned}
F_{\mathrm{x}}(x_2) &= P(\mathrm{x} \leq x_2) \\
&= P[(\mathrm{x} \leq x_1) \cup (x_1 < \mathrm{x} \leq x_2)]
\end{aligned}
$$

Because $\mathrm{x} \leq x_1$ and $x_1 < \mathrm{x} \leq x_2$ are disjoint, we have

$$
\begin{aligned}
F_{\mathrm{x}}(x_2) &= P(\mathrm{x} \leq x_1) + P(x_1 < \mathrm{x} \leq x_2) \\
&= F_{\mathrm{x}}(x_1) + P(x_1 < \mathrm{x} \leq x_2) \qquad (10.25)
\end{aligned}
$$

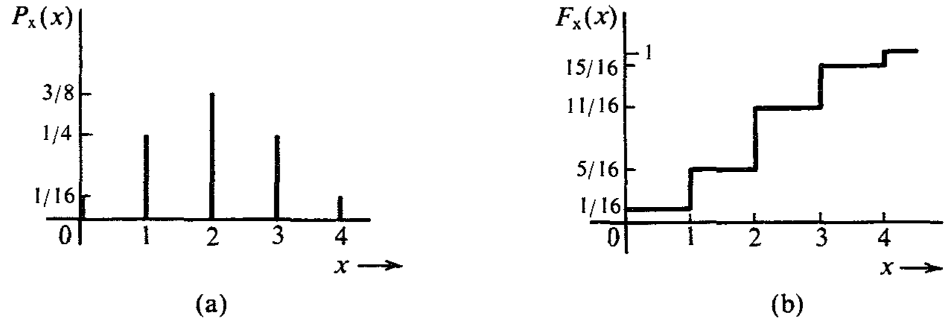
Because $P(x_1 < \mathrm{x} \leq x_2)$ is nonnegative, the result follows.

EXAMPLE 10.12 In a random experiment, a trial consists of four successive tosses of a coin. If we define an RV x as the number of heads appearing in a trial, determine $P_{\mathrm{x}}(x)$ and $F_{\mathrm{x}}(x)$.

A total of 16 distinct equiprobable outcomes are listed in Example 10.4. Various probabilities can be readily determined by counting the outcomes pertaining to a given value of x. For example, only one outcome maps into $x = 0$, whereas six outcomes map into $x = 2$. Hence, $P_{\mathrm{x}}(0) = 1/16$ and $P_{\mathrm{x}}(2) = 6/16$. In the same way, we find

$$P_{\mathrm{x}}(0) = P_{\mathrm{x}}(4) = 1/16$$
$$P_{\mathrm{x}}(1) = P_{\mathrm{x}}(3) = 4/16 = 1/4$$
$$P_{\mathrm{x}}(2) = 6/16 = 3/8$$

The probabilities $P_{\mathrm{x}}(x_i)$ and the corresponding CDF $F_{\mathrm{x}}(x_i)$ are shown in Fig. 10.7.



(a) (b)

Figure 10.7 Probabilities $P_{\mathrm{x}}(x_i)$ and the cumulative distribution function (CDF).

Continuous Random Variables

A continuous RV x can assume any value in a certain interval. In a continuum of any range, an uncountably infinite number of possible values exist, and $P_{\mathrm{x}}(x_i)$, the probability that $\mathrm{x} = x_i$, is one of the uncountably infinite values and is generally zero. Consider the case of a temperature T at a certain location. We may suppose that this temperature can assume any of a range of values. Thus, an uncountably infinite number of possible temperature values may prevail, and the probability that the random variable T assumes a certain value T_i is zero. The situation is somewhat similar to that described in Sec. 3.1.1 in connection with a continuously loaded beam (Fig. 3.5b). There is a loading along the beam at every point, but at any one point the load is zero. The meaningful measure in that case was the loading (or weight) not at a point, but over a finite interval. Similarly, for a continuous RV, the meaningful quantity is not the probability that $\mathrm{x} = x_i$ but the probability that $x < \mathrm{x} \leq x + \Delta x$. For such a measure, the CDF is eminently suited because the latter probability is simply $F_{\mathrm{x}}(x + \Delta x) - F_{\mathrm{x}}(x)$ [see Eq. (10.25)]. Hence, we begin our study of continuous RVs with the CDF.

Properties of the CDF [Eqs. (10.24) and (10.25)] derived earlier are general and are valid for continuous as well as discrete RVs.

Probability Density Function: From Eq. (10.25), we have

$$F_{\mathrm{x}}(x + \Delta x) = F_{\mathrm{x}}(x) + P(x < \mathrm{x} \leq x + \Delta x) \tag{10.26a}$$

If $\Delta x \to 0$, then we can also express $F_{\mathrm{x}}(x + \Delta x)$ via Taylor expansion as

$$F_{\mathrm{x}}(x + \Delta x) \simeq F_{\mathrm{x}}(x) + \frac{dF_{\mathrm{x}}(x)}{dx}\Delta x \tag{10.26b}$$

From Eqs. (10.26), it follows that

$$\lim_{\Delta x \to 0} \frac{dF_{\mathrm{x}}(x)}{dx}\Delta x = P(x < \mathrm{x} \leq x + \Delta x) \tag{10.27}$$

We designated the derivative of $F_{\mathrm{x}}(x)$ with respect to x by $p_{\mathrm{x}}(x)$ (Fig. 10.8),

$$\frac{dF_{\mathrm{x}}(x)}{dx} = p_{\mathrm{x}}(x) \tag{10.28}$$

The function $p_{\mathrm{x}}(x)$ is called the **probability density function (PDF)** of the RV x. It follows from Eq. (10.27) that the probability of observing the RV x in the interval $(x,\ x + \Delta x)$ is $p_{\mathrm{x}}(x)\Delta x$ $(\Delta x \to 0)$. This is the area under the PDF $p_{\mathrm{x}}(x)$ over the interval Δx, as shown in Fig. 10.8b.

From Eq. (10.28) it follows that

$$F_{\mathrm{x}}(x) = \int_{-\infty}^{x} p_{\mathrm{x}}(x)\,dx \tag{10.29}$$

Here we use the fact that $F_{\mathrm{x}}(-\infty) = 0$. We also have from Eq. (10.25)

$$\begin{aligned} P(x_1 < \mathrm{x} \leq x_2) &= F_{\mathrm{x}}(x_2) - F_{\mathrm{x}}(x_1) \\ &= \int_{-\infty}^{x_2} p_{\mathrm{x}}(x)\,dx - \int_{-\infty}^{x_1} p_{\mathrm{x}}(x)\,dx \\ &= \int_{x_1}^{x_2} p_{\mathrm{x}}(x)\,dx \end{aligned} \tag{10.30}$$

Thus, the probability of observing x in any interval $(x_1,\ x_2)$ is given by the area under the PDF $p_{\mathrm{x}}(x)$ over the interval $(x_1,\ x_2)$, as shown in Fig. 10.8. Compare this with a continuously loaded beam (Fig. 3.5), where the weight over any interval was given by an integral of the loading density over the interval.

Because $F_{\mathrm{x}}(\infty) = 1$, we have

$$\int_{-\infty}^{\infty} p_{\mathrm{x}}(x)\,dx = 1 \tag{10.31}$$

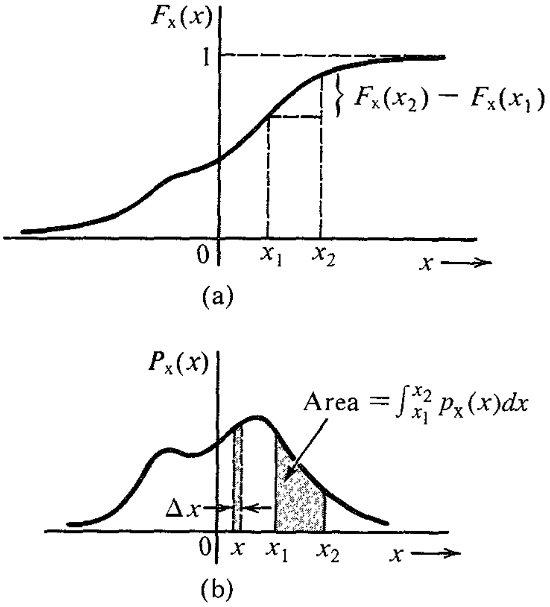


Figure 10.8 Cumulative distribution function (CDF) and probability density function (PDF).

This also follows from the fact that the integral in Eq. (10.31) represents the probability of observing x in the interval $(-\infty, \infty)$. Every PDF must satisfy the condition in Eq. (10.31). It is also evident that the PDF must be nonnegative, that is,

$$p_{\mathrm{x}}(x) \geq 0$$

Although it is true that the probability of an impossible event is **0** and that of a certain event is **1**, the converse is not true. An event whose probability is **0** is not necessarily an impossible event, and an event with a probability of **1** is not necessarily a certain event. This may be illustrated by the following example. The temperature T of a certain city on a summer day is an RV taking on any value in the range of 5 to 50°C. Because the PDF $p_{\mathrm{T}}(T)$ is continuous, the probability that T = 34.56, for example, is zero. But this is not an impossible event. Similarly, the probability that T takes on any value but 34.56 is **1**, although this is not a certain event. In fact, a continuous RV x takes every value in a certain range. Yet $p_{\mathrm{x}}(x)$, the probability that $\mathrm{x} = x$, is zero for every x in that range.

We can also determine the PDF $p_{\mathrm{x}}(x)$ for a discrete random variable. Because the CDF $F_{\mathrm{x}}(x)$ for the discrete case is always a sequence of step functions (Fig. 10.7), the PDF (the derivative of the CDF) will consist of a train of impulses. If an RV x takes values $x_1, x_2, \ldots, x_n$ with probabilities $a_1, a_2, \ldots, a_n$, respectively, then

$$F_{\mathrm{x}}(x) = a_1 u(x - x_1) + a_2 u(x - x_2) + \cdots + a_n u(x - x_n) \tag{10.32a}$$

This can be easily verified from Example 10.12 (Fig. 10.7). Hence,

$$\begin{aligned} p_{\mathrm{x}}(x) &= a_1\,\delta(x - x_1) + a_2\,\delta(x - x_2) + \cdots + a_n\,\delta(x - x_n) \\ &= \sum_{r=1}^{n} a_r\,\delta(x - x_r) \end{aligned} \tag{10.32b}$$

It is, of course, possible to have a mixed case, where a PDF may have a continuous part and an impulsive part (see Prob. 10.2-4).

The Gaussian PDF: Consider a PDF (Fig. 10.9a)

$$p_{\mathrm{x}}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{10.33}$$

This is a case of the well-known **gaussian**, or **normal**, probability density.

The CDF $F_{\mathrm{x}}(x)$ in this case is

$$F_{\mathrm{x}}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2}\,dx$$

This integral cannot be evaluated in a closed form and must be computed numerically. It is convenient to use the function $Q(.)$, defined as[2]

$$Q(y) = \frac{1}{\sqrt{2\pi}} \int_{y}^{\infty} e^{-x^2/2}\,dx \tag{10.34}$$

The area under $p_{\mathrm{x}}(x)$ from y to ∞ (shown shaded in Fig. 10.9a) is* $Q(y)$. From the symmetry of $p_{\mathrm{x}}(x)$ about the origin, and the fact that the total area under $p_{\mathrm{x}}(x) = 1$, it follows that

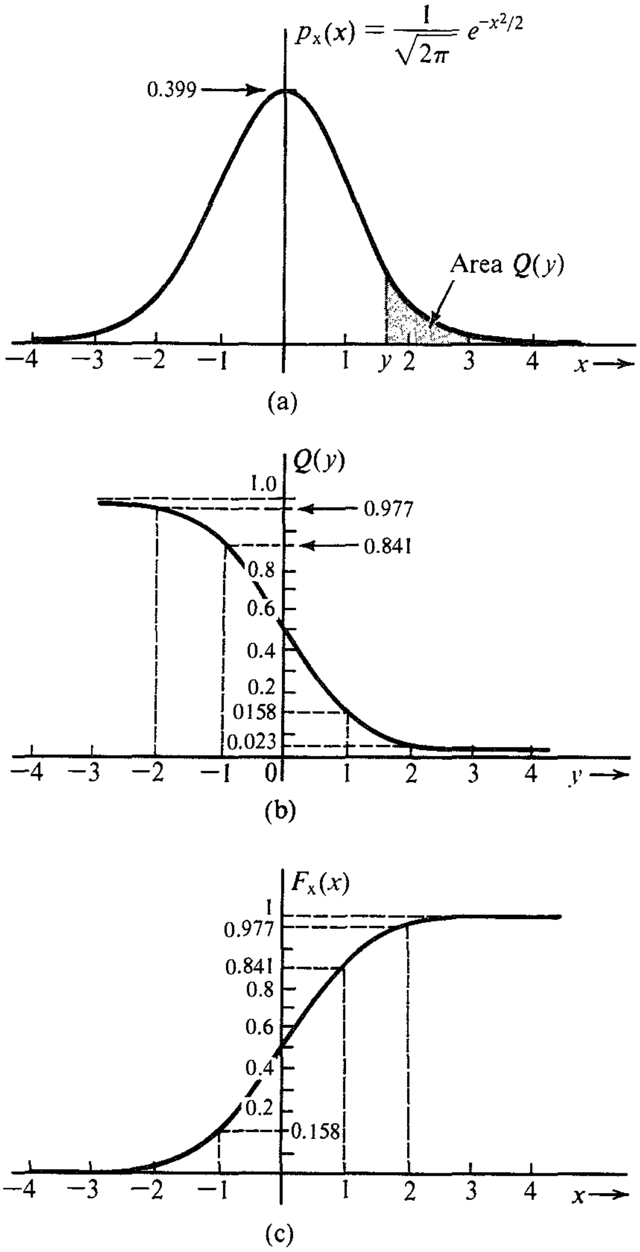


Figure 10.9 (a) Gaussian PDF. (b) Function $Q(y)$. (c) CDF of the gaussian PDF.

---

* The function $Q(x)$ is closely related to functions erf (x) and erfc (x),

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-y^2}\, dy = 2Q(x\sqrt{2})$$

Therefore,

$$Q(x) = \frac{1}{2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) = \frac{1}{2}\left[1 - \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

Table 10.2
$Q(x)$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0000	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1000	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2000	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3000	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4000	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5000	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6000	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7000	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8000	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9000	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.000	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.100	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.200	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.9853E-01
1.300	.9680E-01	.9510E-01	.9342E-01	.9176E-01	.9012E-01	.8851E-01	.8691E-01	.8534E-01	.8379E-01	.8226E-01
1.400	.8076E-01	.7927E-01	.7780E-01	.7636E-01	.7493E-01	.7353E-01	.7215E-01	.7078E-01	.6944E-01	.6811E-01
1.500	.6681E-01	.6552E-01	.6426E-01	.6301E-01	.6178E-01	.6057E-01	.5938E-01	.5821E-01	.5705E-01	.5592E-01
1.600	.5480E-01	.5370E-01	.5262E-01	.5155E-01	.5050E-01	.4947E-01	.4846E-01	.4746E-01	.4648E-01	.4551E-01
1.700	.4457E-01	.4363E-01	.4272E-01	.4182E-01	.4093E-01	.4006E-01	.3920E-01	.3836E-01	.3754E-01	.3673E-01
1.800	.3593E-01	.3515E-01	.3438E-01	.3362E-01	.3288E-01	.3216E-01	.3144E-01	.3074E-01	.3005E-01	.2938E-01
1.900	.2872E-01	.2807E-01	.2743E-01	.2680E-01	.2619E-01	.2559E-01	.2500E-01	.2442E-01	.2385E-01	.2330E-01
2.000	.2275E-01	.2222E-01	.2169E-01	.2118E-01	.2068E-01	.2018E-01	.1970E-01	.1923E-01	.1876E-01	.1831E-01
2.100	.1786E-01	.1743E-01	.1700E-01	.1659E-01	.1618E-01	.1578E-01	.1539E-01	.1500E-01	.1463E-01	.1426E-01
2.200	.1390E-01	.1355E-01	.1321E-01	.1287E-01	.1255E-01	.1222E-01	.1191E-01	.1160E-01	.1130E-01	.1101E-01
2.300	.1072E-01	.1044E-01	.1017E-01	.9903E-02	.9642E-02	.9387E-02	.9137E-02	.8894E-02	.8656E-02	.8424E-02
2.400	.8198E-02	.7976E-02	.7760E-02	.7549E-02	.7344E-02	.7143E-02	.6947E-02	.6756E-02	.6569E-02	.6387E-02
2.500	.6210E-02	.6037E-02	.5868E-02	.5703E-02	.5543E-02	.5386E-02	.5234E-02	.5085E-02	.4940E-02	.4799E-02
2.600	.4661E-02	.4527E-02	.4396E-02	.4269E-02	.4145E-02	.4025E-02	.3907E-02	.3793E-02	.3681E-02	.3573E-02
2.700	.3467E-02	.3364E-02	.3264E-02	.3167E-02	.3072E-02	.2980E-02	.2890E-02	.2803E-02	.2718E-02	.2635E-02
2.800	.2555E-02	.2477E-02	.2401E-02	.2327E-02	.2256E-02	.2186E-02	.2118E-02	.2052E-02	.1988E-02	.1926E-02
2.900	.1866E-02	.1807E-02	.1750E-02	.1695E-02	.1641E-02	.1589E-02	.1538E-02	.1489E-02	.1441E-02	.1395E-02
3.000	.1350E-02	.1306E-02	.1264E-02	.1223E-02	.1183E-02	.1144E-02	.1107E-02	.1070E-02	.1035E-02	.1001E-02
3.100	.9676E-03	.9354E-03	.9043E-03	.8740E-03	.8447E-03	.8164E-03	.7888E-03	.7622E-03	.7364E-03	.7114E-03
3.200	.6871E-03	.6637E-03	.6410E-03	.6190E-03	.5976E-03	.5770E-03	.5571E-03	.5377E-03	.5190E-03	.5009E-03
3.300	.4834E-03	.4665E-03	.4501E-03	.4342E-03	.4189E-03	.4041E-03	.3897E-03	.3758E-03	.3624E-03	.3495E-03
3.400	.3369E-03	.3248E-03	.3131E-03	.3018E-03	.2909E-03	.2802E-03	.2701E-03	.2602E-03	.2507E-03	.2415E-03
3.500	.2326E-03	.2241E-03	.2158E-03	.2078E-03	.2001E-03	.1926E-03	.1854E-03	.1785E-03	.1718E-03	.1653E-03
3.600	.1591E-03	.1531E-03	.1473E-03	.1417E-03	.1363E-03	.1311E-03	.1261E-03	.1213E-03	.1166E-03	.1121E-03
3.700	.1078E-03	.1036E-03	.9961E-04	.9574E-04	.9201E-04	.8842E-04	.8496E-04	.8162E-04	.7841E-04	.7532E-04
3.800	.7235E-04	.6948E-04	.6673E-04	.6407E-04	.6152E-04	.5906E-04	.5669E-04	.5442E-04	.5223E-04	.5012E-04
3.900	.4810E-04	.4615E-04	.4427E-04	.4247E-04	.4074E-04	.3908E-04	.3747E-04	.3594E-04	.3446E-04	.3304E-04
4.000	.3167E-04	.3036E-04	.2910E-04	.2789E-04	.2673E-04	.2561E-04	.2454E-04	.2351E-04	.2252E-04	.2157E-04
4.100	.2066E-04	.1978E-04	.1894E-04	.1814E-04	.1737E-04	.1662E-04	.1591E-04	.1523E-04	.1458E-04	.1395E-04
4.200	.1335E-04	.1277E-04	.1222E-04	.1168E-04	.1118E-04	.1069E-04	.1022E-04	.9774E-05	.9345E-05	.8934E-05
4.300	.8540E-05	.8163E-05	.7801E-05	.7455E-05	.7124E-05	.8807E-05	.6503E-05	.6212E-05	.5934E-05	.5668E-05
4.400	.5413E-05	.5169E-05	.4935E-05	.4712E-05	.4498E-05	.4294E-05	.4098E-05	.3911E-05	.3732E-05	.3561E-05
4.500	.3398E-05	.3241E-05	.3092E-05	.2949E-05	.2813E-05	.2682E-05	.2558E-05	.2439E-05	.2325E-05	.2216E-05
4.600	.2112E-05	.2013E-05	.1919E-05	.1828E-05	.1742E-05	.1660E-05	.1581E-05	.1506E-05	.1434E-05	.1366E-05
4.700	.1301E-05	.1239E-05	.1179E-05	.1123E-05	.1069E-05	.1017E-05	.9680E-06	.9211E-06	.8765E-06	.8339E-06
4.800	.7933E-06	.7547E-06	.7178E-06	.6827E-06	.6492E-06	.6173E-06	.5869E-06	.5580E-06	.5304E-06	.5042E-06
4.900	.4792E-06	.4554E-06	.4327E-06	.4111E-06	.3906E-06	.3711E-06	.3525E-06	.3448E-06	.3179E-06	.3019E-06
5.000	.2867E-06	.2722E-06	.2584E-06	.2452E-06	.2328E-06	.2209E-06	.2096E-06	.1989E-06	.1887E-06	.1790E-06
5.100	.1698E-06	.1611E-06	.1528E-06	.1449E-06	.1374E-06	.1302E-06	.1235E-06	.1170E-06	.1109E-06	.1051E-06

(continued)

Table 10.2
Continued

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
5.200	.9964E-07	.9442E-07	.8946E-07	.8476E-07	.8029E-07	.7605E-07	.7203E-07	.6821E-07	.6459E-07	.6116E-07
5.300	.5790E-07	.5481E-07	.5188E-07	.4911E-07	.4647E-07	.4398E-07	.4161E-07	.3937E-07	.3724E-07	.3523E-07
5.400	.3332E-07	.3151E-07	.2980E-07	.2818E-07	.2664E-07	.2518E-07	.2381E-07	.2250E-07	.2127E-07	.2010E-07
5.500	.1899E-07	.1794E-07	.1695E-07	.1601E-07	.1512E-07	.1428E-07	.1349E-07	.1274E-07	.1203E-07	.1135E-07
5.600	.1072E-07	.1012E-07	.9548E-08	.9010E-08	.8503E-08	.8022E-08	.7569E-08	.7140E-08	.6735E-08	.6352E-08
5.700	.5990E-08	.5649E-08	.5326E-08	.5022E-08	.4734E-08	.4462E-08	.4206E-08	.3964E-08	.3735E-08	.3519E-08
5.800	13316E-08	.3124E-08	.2942E-08	.2771E-08	.2610E-08	.2458E-08	.2314E-08	.2179E-08	.2051E-08	.1931E-08
5.900	.1818E-08	.1711E-08	.1610E-08	.1515E-08	.1425E-08	.1341E-08	.1261E-08	.1186E-08	.1116E-08	.1049E-08
6.000	.9866E-09	.9276E-09	.8721E-09	.8198E-09	.7706E-09	.7242E-09	.6806E-09	.6396E-09	.6009E-09	.5646E-09
6.100	.5303E-09	.4982E-09	.4679E-09	.4394E-09	.4126E-09	.3874E-09	.3637E-09	.3414E-09	.3205E-09	.3008E-09
6.200	.2823E-09	.2649E-09	.2486E-09	.2332E-09	.2188E-09	.2052E-09	.1925E-09	.1805E-09	.1692E-09	.1587E-09
6.300	.1488E-09	.1395E-09	.1308E-09	.1226E-09	.1149E-09	.1077E-09	.1009E-09	.9451E-10	.8854E-10	.8294E-10
6.400	.7769E-10	.7276E-10	.6814E-10	.6380E-10	.5974E-10	.5593E-10	.5235E-10	.4900E-10	.4586E-10	.4292E-10
6.500	.4016E-10	.3758E-10	.3515E-10	.3288E-10	.30767E-10	.2877E-10	.2690E-10	.2516E-10	.2352E-10	.2199E-10
6.600	.2056E-10	.1922E-10	.1796E-10	.1678E-10	.1568E-10	.1465E-10	.1369E-10	.1279E-10	.1195E-10	.1116E-10
6.700	.1042E-10	.9731E-11	.9086E-11	.8483E-11	.7919E-11	.7392E-11	.6900E-11	.6439E-11	.6009E-11	.5607E-11
6.800	.5231E-11	.4880E-11	.4552E-11	.4246E-11	.3960E-11	.3692E-11	.3443E-11	.3210E-11	.2993E-11	.2790E-11
6.900	.2600E-11	.2423E-11	.2258E-11	.2104E-11	.1960E-11	.1826E-11	.1701E-11	.1585E-11	.1476E-11	.1374E-11
7.000	.1280E-11	.1192E-11	.1109E-11	.1033E-11	.9612E-12	.8946E-12	.8325E-12	.7747E-12	.7208E-12	.6706E-12
7.100	.6238E-12	.5802E-12	.5396E-12	.5018E-12	.4667E-12	.4339E-12	.4034E-12	.3750E-12	.3486E-12	.3240E-12
7.200	.3011E-12	.2798E-12	.2599E-12	.2415E-12	.2243E-12	.2084E-12	.1935E-12	.1797E-12	.1669E-12	.1550E-12
7.300	.1439E-12	.1336E-12	.1240E-12	.1151E-12	.1068E-12	.9910E-13	.9196E-13	.8531E-13	.7914E-13	.7341E-13
7.400	.6809E-13	.6315E-13	.5856E-13	.5430E-13	.5034E-13	.4667E-13	.4326E-13	.4010E-13	.3716E-13	.3444E-13
7.500	.3191E-13	.2956E-13	.2739E-13	.2537E-13	.2350E-13	.2176E-13	.2015E-13	.1866E-13	.1728E-13	.1600E-13
7.600	.1481E-13	.1370E-13	.1268E-13	.1174E-13	.1086E-13	.1005E-13	.9297E-14	.8600E-14	.7954E-14	.7357E-14
7.700	.6803E-14	.6291E-14	.5816E-14	.5377E-14	.4971E-14	.4595E-14	.4246E-14	.3924E-14	.3626E-14	.3350E-14
7.800	.3095E-14	.2859E-14	.2641E-14	.2439E-14	.2253E-14	.2080E-14	.1921E-14	.1773E-14	.1637E-14	.1511E-14
7.900	.1395E-14	.1287E-14	.1188E-14	.1096E-14	.1011E-14	.9326E-15	.8602E-15	.7934E-15	.7317E-15	.6747E-15
8.000	.6221E-15	.5735E-15	.5287E-15	.4874E-15	.4492E-15	.4140E-15	.3815E-15	.3515E-15	.3238E-15	.2983E-15
8.100	.2748E-15	.2531E-15	.2331E-15	.2146E-15	.1976E-15	.1820E-15	.1675E-15	.1542E-15	.1419E-15	.1306E-15
8.200	.1202E-15	.1106E-15	.1018E-15	.9361E-16	.8611E-16	.7920E-16	.7284E-16	.6698E-16	.6159E-16	.5662E-16
8.300	.5206E-16	.4785E-16	.4398E-16	.4042E-16	.3715E-16	.3413E-16	.3136E-16	.2881E-16	.2646E-16	.2431E-16
8.400	.2232E-16	.2050E-16	.1882E-16	.1728E-16	.1587E-16	.1457E-16	.1337E-16	.1227E-16	.1126E-16	.1033E-16
8.500	.9480E-17	.8697E-17	.7978E-17	.7317E-17	.6711E-17	.6154E-17	.5643E-17	.5174E-17	.4744E-17	.4348E-17
8.600	.3986E-17	.3653E-17	.3348E-17	.3068E-17	.2811E-17	.2575E-17	.2359E-17	.2161E-17	.1979E-17	.1812E-17
8.700	.1659E-17	.1519E-17	.1391E-17	.1273E-17	.1166E-17	.1067E-17	.9763E-18	.8933E-18	.8174E-18	.7478E-18
8.800	.6841E-18	.6257E-18	.5723E-18	.5234E-18	.4786E-18	.4376E-18	.4001E-18	.3657E-18	.3343E-18	.3055E-18
8.900	.2792E-18	.2552E-18	.2331E-18	.2130E-18	.1946E-18	.1777E-18	.1623E-18	.1483E-18	.1354E-18	.1236E-18
9.000	.1129E-18	.1030E-18	.9404E-19	.8584E-19	.7834E-19	.7148E-19	.6523E-19	.5951E-19	.5429E-19	.4952E-19
9.100	.4517E-19	.4119E-19	.3756E-19	.3425E-19	.3123E-19	.2847E-19	.2595E-19	.2365E-19	.2155E-19	.1964E-19
9.200	.1790E-19	.1631E-19	.1486E-19	.1353E-19	.1232E-19	.1122E-19	.1022E-19	.9307E-20	.8474E-20	.7714E-20
9.300	.7022E-20	.6392E-20	.5817E-20	.5294E-20	.4817E-20	.4382E-20	.3987E-20	.3627E-20	.3299E-20	.3000E-20
9.400	.2728E-20	.2481E-20	.2255E-20	.2050E-20	.1864E-20	.1694E-20	.1540E-20	.1399E-20	.1271E-20	.1155E-20
9.500	.1049E-20	.9533E-21	.8659E-21	.7864E-21	.7142E-21	.6485E-21	.5888E-21	.5345E-21	.4852E-21	.4404E-21
9.600	.3997E-21	.3627E-21	.3292E-21	.2986E-21	.2709E-21	.2458E-21	.2229E-21	.2022E-21	.1834E-21	.1663E-21
9.700	.1507E-21	.1367E-21	.1239E-21	.1123E-21	.1018E-21	.9223E-22	.8358E-22	.7573E-22	.6861E-22	.6215E-22
9.800	.5629E-22	.5098E-22	.4617E-22	.4181E-22	.3786E-22	.3427E-22	.3102E-22	.2808E-22	.2542E-22	.2300E-22
9.900	.2081E-22	.1883E-22	.1704E-22	.1541E-22	.1394E-22	.1261E-22	.1140E-22	.1031E-22	.9323E-23	.8429E-23
10.00	.7620E-23	.6888E-23	.6225E-23	.5626E-23	.5084E-23	.4593E-23	.4150E-23	.3749E-23	.3386E-23	.3058E-23

Notes: (1) E-01 should be read as $\times 10^{-1}$; E-02 should be read as $\times 10^{-2}$, and so on.

(2) This table lists $Q(x)$ for x in the range of 0 to 10 in the increments of 0.01. To find $Q(5.36)$, for example, look up the row starting with $x = 5.3$. The sixth entry in this row (under 0.06) is the desired value 0.4161×10^{-7}.

$$Q(-y) = 1 - Q(y) \tag{10.35}$$

The function $Q(x)$ is tabulated in Table 10.2 (see also Fig. 10.11d). This function is widely tabulated and can be found in most of the standard mathematical tables.[2,3] It can be shown that[4]

$$Q(x) \simeq \frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \qquad \text{for } x \gg 1 \tag{10.36a}$$

For example, when $x = 2$, the error in this approximation is 18.7%. But for $x = 4$ it is 10.4% and for $x = 6$ it is 2.3%.

A much better approximation to $Q(x)$ is

$$Q(x) \simeq \frac{1}{x\sqrt{2\pi}} \left(1 - \frac{0.7}{x^2}\right) e^{-x^2/2} \qquad x > 2 \tag{10.36b}$$

The error in this approximation is just within 1% for $x > 2.15$. For larger values of x the error approaches 0.

Observe that for the PDF in Fig. 10.9a, the CDF is given by (Fig. 10.9c)

$$F_{\mathrm{x}}(x) = 1 - Q(x) \tag{10.37}$$

A more general gaussian density function is (Fig. 10.10)

$$p_{\mathrm{x}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \tag{10.38}$$

For this case,

$$F_{\mathrm{x}}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-m)^2/2\sigma^2}\, dx$$

Letting $(x - m)/\sigma = z$,

$$\begin{aligned} F_{\mathrm{x}}(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-m)/\sigma} e^{-z^2/2}\, dz \\ &= 1 - Q\left(\frac{x-m}{\sigma}\right) \end{aligned} \tag{10.39a}$$

Therefore,

$$P(\mathrm{x} \le x) = 1 - Q\left(\frac{x-m}{\sigma}\right) \tag{10.39b}$$

and

$$P(\mathrm{x} > x) = Q\left(\frac{x-m}{\sigma}\right) \tag{10.39c}$$

The gaussian PDF is perhaps the most important PDF in the area of communications. The majority of the noise processes observed in practice are gaussian. The amplitude n of a gaussian noise signal is an RV with a gaussian PDF. This means the probability of observing n in an interval $(n,\ n + \Delta n)$ is $p_{\mathrm{n}}(n)\Delta n$, where $p_{\mathrm{n}}(n)$ is of the form in Eq. (10.38) [with $m = 0$].

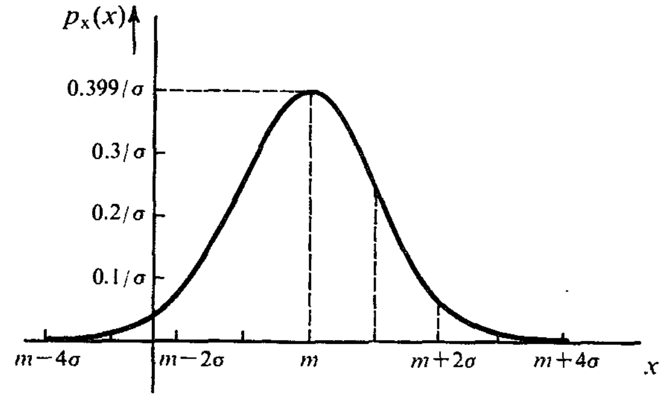


Figure 10.10 Gaussian PDF.

EXAMPLE 10.13 Threshold Detection

Over a certain binary channel, messages $m = 0$ and **1** are transmitted with equal probability using a positive and a negative pulse, respectively. The received pulse corresponding to **1** is $p(t)$, shown in Fig. 10.11a, and the received pulse corresponding to **0** is $-p(t)$. Let the peak amplitude of $p(t)$ be A_p at $t = T_p$. Because of the channel noise $\mathrm{n}(t)$, the received pulses will be $\pm p(t) + \mathrm{n}(t)$ (Fig. 10.11b). To detect the pulses at the receiver, each pulse is sampled at its peak amplitude. In the absence of noise, the sampler output is either A_p (for $m = 1$) or $-A_p$ (for $m = 0$). Because of the channel noise, the sampler output is $\pm A_p + \mathrm{n}$, where n, the noise amplitude at the sampling instant (Fig. 10.11b), is an RV. For gaussian noise, the PDF of n is (Fig. 10.11c)

$$p_{\mathrm{n}}(n) = \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} e^{-n^2/2\sigma_{\mathrm{n}}^2} \tag{10.40}$$

Because of the symmetry of the situation, the optimum detection threshold is zero; that is, the received pulse is detected as a **1** or a **0**, depending on whether the sample value is positive or negtive.

Because noise amplitudes range from $-\infty$ to ∞, the sample value $-A_p + \mathrm{n}$ can occasionally be positive, causing the received **0** to be read as **1** (see Fig. 10.11b). Similarly, $A_p + \mathrm{n}$ can occasionally be negative, causing the received **1** to be read as **0**. If **0** is transmitted, it will be detected as **1** if $-A_p + \mathrm{n} > 0$, that is, if $\mathrm{n} > A_p$.

If $P(\epsilon|\mathbf{0})$ is the error probability given that **0** is transmitted, then

$$P(\epsilon|\mathbf{0}) = P(\mathrm{n} > A_p)$$

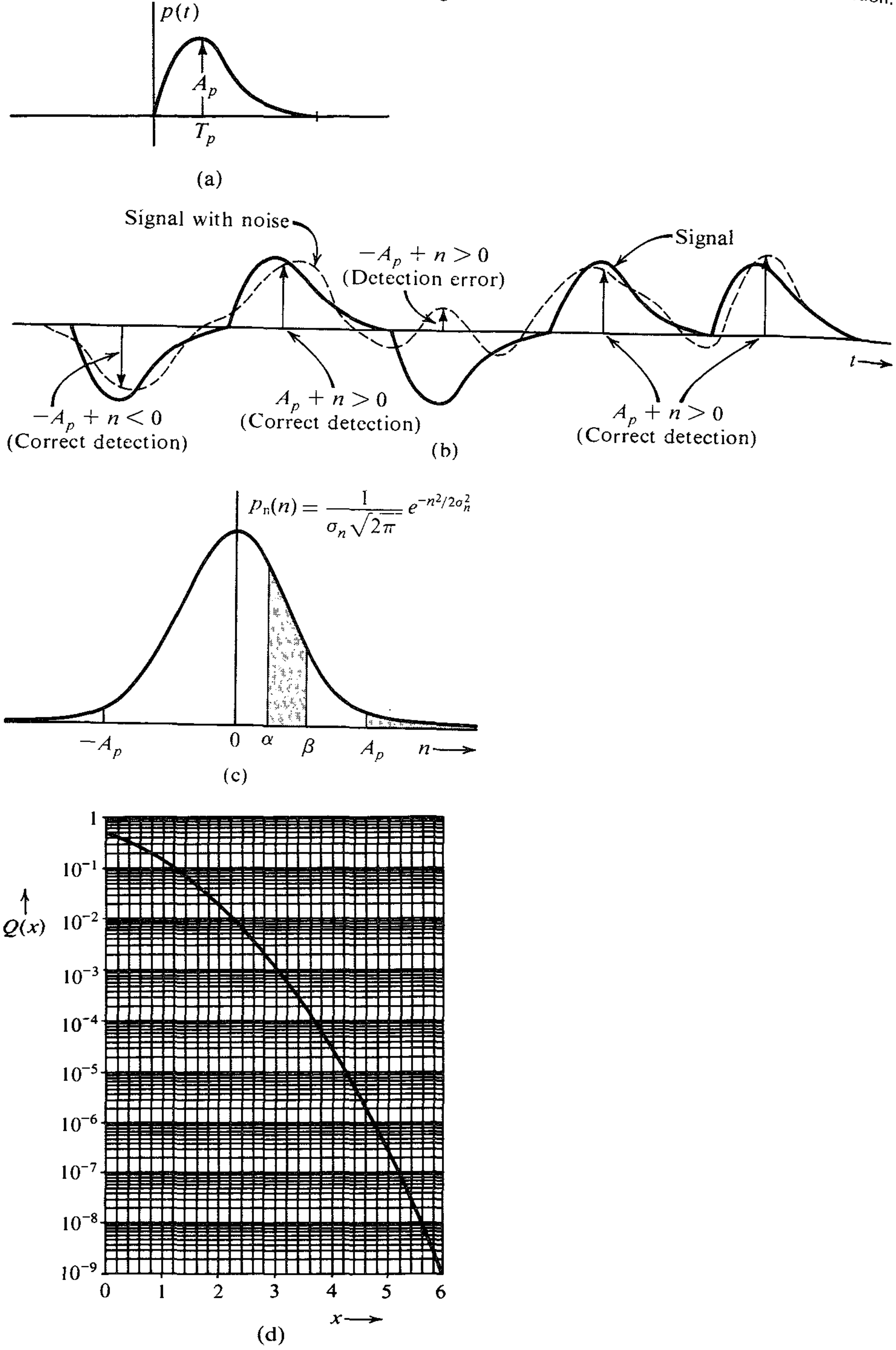
Because $P(\mathrm{n} > A_p)$ is the shaded area in Fig. 10.11c, from Eq. (10.39c) [with $m = 0$] it follows that

$$P(\epsilon|\mathbf{0}) = Q\left(\frac{A_p}{\sigma_{\mathrm{n}}}\right) \tag{10.41a}$$

Similarly,

$$\begin{aligned} P(\epsilon|\mathbf{1}) &= P(\mathrm{n} < -A_p) \\ &= Q\left(\frac{A_p}{\sigma_{\mathrm{n}}}\right) = P(\epsilon|\mathbf{0}) \end{aligned} \tag{10.41b}$$

Figure 10.11 Error probability in threshold detection.



and

$$
\begin{aligned}
P_e &= \sum_i P(\epsilon,\ m_i) \\
&= \sum_i P(m_i)P(\epsilon|m_i) \\
&= Q\left(\frac{A_p}{\sigma_n}\right)\sum_i P(m_i) \\
&= Q\left(\frac{A_p}{\sigma_{\rm n}}\right)
\end{aligned}
\tag{10.41c}
$$

The error probability P_e can be found from Fig. 10.11d.

Joint Distribution: For two RVs x and y, we define a CDF $F_{\rm xy}(x,\ y)$ as follows:

$$P(\mathrm{x} \le x \text{ and } \mathrm{y} \le y) = F_{\rm xy}(x,\ y) \tag{10.42}$$

and the joint PDF $p_{\rm xy}(x,\ y)$ as

$$p_{\rm xy}(x,\ y) = \frac{\partial^2}{\partial x\,\partial y}F_{\rm xy}(x,\ y) \tag{10.43}$$

Arguing along lines similar to those used for a single variable, we can show that

$$\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}} p_{\rm xy}(x,\ y)\Delta x\,\Delta y = P(x < \mathrm{x} \le x+\Delta x,\ y < \mathrm{y} \le y+\Delta y) \tag{10.44}$$

Hence, the probability of observing the variables x in the interval $(x,\ x+\Delta x)$ and y in the interval $(y,\ y+\Delta y)$ jointly is given by the volume under the joint PDF $p_{\rm xy}(x,\ y)$ over the region bounded by $(x,\ x+\Delta x)$ and $(y,\ y+\Delta y)$, as shown in Fig. 10.12a.

From Eq. (10.44), it follows that

$$P(x_1 < \mathrm{x} \le x_2,\ y_1 < \mathrm{y} \le y_2) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} p_{\rm xy}(x,\ y)\,dx\,dy \tag{10.45}$$

Thus, the probability of jointly observing x in the interval (x_1, x_2) and y in the interval $(y_1,\ y_2)$ is the volume under the PDF over the region bounded by $(x_1,\ x_2)$ and $(y_1,\ y_2)$.

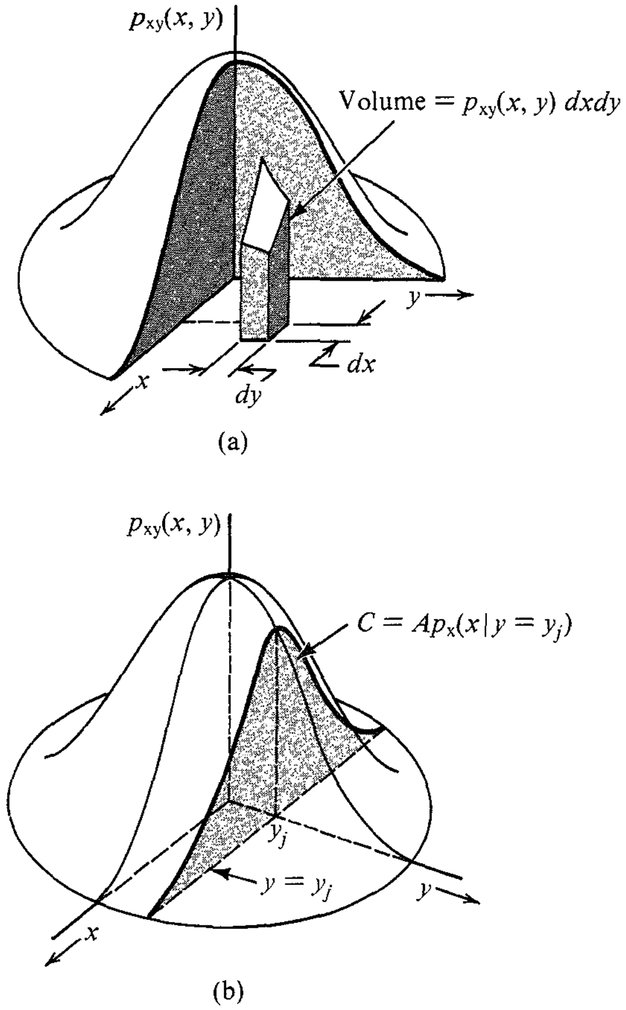
The event of observing x in the interval $(-\infty,\ \infty)$ and observing y in the interval $(-\infty,\ \infty)$ is a certainty. Hence,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\rm xy}(x,\ y)\,dx\,dy = 1 \tag{10.46}$$

Thus, the total volume under the joint PDF must be unity.

When we are dealing with two RVs x and y, the individual probability densities $p_{\rm x}(x)$ and $p_{\rm y}(y)$ can be obtained from the joint density $p_{\rm xy}(x,\ y)$. These individual densities are also called **marginal densities**. To obtain these densities, we note that $p_{\rm x}(x)\Delta x$ is the probability of observing x in the interval $(x,\ x+\Delta x)$. The value of y may lie anywhere in the interval $(-\infty,\ \infty)$. Hence,

Figure 10.12 (a) Joint PDF. (b) Conditional PDF.


$p_{xy}(x, y)$
Volume = $p_{xy}(x, y)\,dxdy$
y
dx
x
dy
(a)
$p_{xy}(x, y)$
$C = Ap_x(x\,|\,y = y_j)$
y_j
$y = y_j$
y
x
(b)

$$\lim_{\Delta x \to 0} p_x(x)\Delta x = \text{probability } (x < \text{x} \le x + \Delta x, \ -\infty < \text{y} \le \infty)$$

$$= \lim_{\Delta x \to 0} \int_x^{x+\Delta x} \int_{-\infty}^{\infty} p_{xy}(x, \ y)\, dx\, dy$$

$$= \lim_{\Delta x \to 0} \int_{-\infty}^{\infty} p_{xy}(x, \ y)\, dy \int_x^{x+\Delta x} dx$$

$$= \lim_{\Delta x \to 0} \Delta x \int_{-\infty}^{\infty} p_{xy}(x, \ y)\, dy$$

The last two steps follow from the fact that $p_{xy}(x, \ y)$ is constant over $(x, \ x + \Delta x)$ because $\Delta x \to 0$. Therefore,

$$p_x(x) = \int_{-\infty}^{\infty} p_{xy}(x, \ y)\, dy \tag{10.47a}$$

Similarly,

$$p_{\mathrm{y}}(y) = \int_{-\infty}^{\infty} p_{\mathrm{xy}}(x, y)\, dx \tag{10.47b}$$

These results may be generalized for n RVs $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$.

Conditional Densities: The concept of conditional probabilities can be extended to the case of continuous RVs. We define the conditional PDF $p_{\mathrm{x}|\mathrm{y}}(x|y_j)$ as the PDF of x given that y has a value y_j. This is equivalent to saying that $p_{\mathrm{x}|\mathrm{y}}(x|y_j)\Delta x$ is the probability of observing x in the range $(x, x + \Delta x)$, given that $\mathrm{y} = y_j$. The probability density $p_{\mathrm{x}|\mathrm{y}}(x|y_j)$ is the intersection of the plane $\mathrm{y} = y_j$ with the joint PDF $p_{\mathrm{xy}}(x, y)$ (Fig. 10.12b). Because every PDF must have unit area, however, we must normalize the area under the intersection curve C to unity to get the desired PDF. Hence, C is $Ap_{\mathrm{x}|\mathrm{y}}(x|y)$, where A is the area under C. An extension of the results derived for the discrete case yields

$$p_{\mathrm{x}|\mathrm{y}}(x|y)p_{\mathrm{y}}(y) = p_{\mathrm{xy}}(x, y) \tag{10.48a}$$

$$p_{\mathrm{y}|\mathrm{x}}(y|x)p_{\mathrm{x}}(x) = p_{\mathrm{xy}}(x, y) \tag{10.48b}$$

and

$$p_{\mathrm{x}|\mathrm{y}}(x|y) = \frac{p_{\mathrm{y}|\mathrm{x}}(y|x)p_{\mathrm{x}}(x)}{p_{\mathrm{y}}(y)} \tag{10.49a}$$

Equation (10.49a) is Bayes' rule for continuous RVs. When we have mixed variables (i.e., discrete and continuous), the mixed form of Bayes' rule is

$$P_{\mathrm{x}|\mathrm{y}}(x|y)p_{\mathrm{y}}(y) = P_{\mathrm{x}}(x)p_{\mathrm{y}|\mathrm{x}}(y|x) \tag{10.49b}$$

where x is a discrete RV and y is a continuous RV

Continuous RVs x and y are said to be independent if

$$p_{\mathrm{x}|\mathrm{y}}(x|y) = p_{\mathrm{x}}(x) \tag{10.50a}$$

In this case from Eqs. (10.50a) and (10.49) it follows that

$$p_{\mathrm{y}|\mathrm{x}}(y|x) = p_{\mathrm{y}}(y) \tag{10.50b}$$

This implies that for independent RVs x and y,

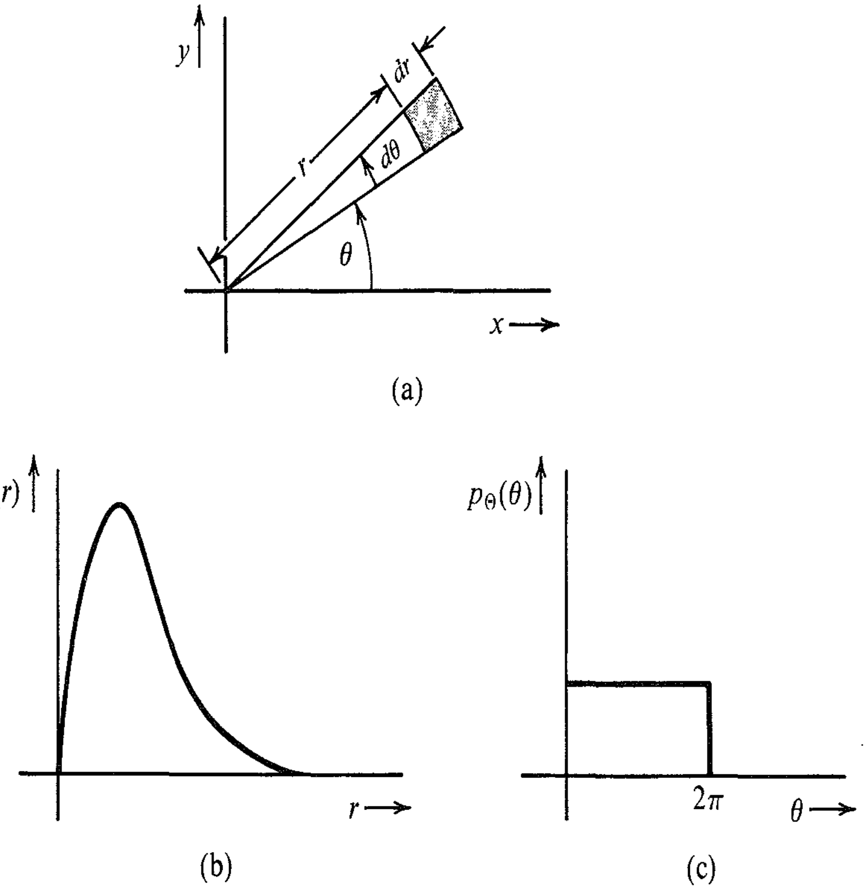
$$p_{\mathrm{xy}}(x, y) = p_{\mathrm{x}}(x)p_{\mathrm{y}}(y) \tag{10.50c}$$

EXAMPLE 10.14 Rayleigh Density

The Rayleigh density is characterized by the PDF (Fig. 10.13b)

$$p_{\mathrm{r}}(r) = \begin{cases} \dfrac{r}{\sigma^2} e^{-r^2/2\sigma^2} & r \geq 0 \\ 0 & r < 0 \end{cases} \tag{10.51}$$

A Rayleigh RV can be derived from two independent gaussian RVs as follows. Let x and y be independent gaussian variables with identical PDFs:


(a)

(b) (c)

Figure 10.13 Derivation of Rayleigh density.

$$p_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

$$p_y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/2\sigma^2}$$

Then

$$p_{xy}(x,\ y) = p_x(x)p_y(y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{10.52}$$

The joint density appears somewhat like the bell-shaped surface shown in Fig. 10.12. The points in the x, y plane can also be described in polar coordinates as $(r,\ \theta)$, where (Fig. 10.13a)

$$\mathrm{r} = \sqrt{\mathrm{x}^2 + \mathrm{y}^2} \qquad \Theta = \tan^{-1}\frac{\mathrm{y}}{\mathrm{x}}$$

In Fig. 10.13a, the shaded region represents $r < \mathrm{r} \leq r + dr$ and $\theta < \Theta \leq \theta + d\theta$ (where dr and $d\theta$ both $\to 0$). Hence, if $p_{\mathrm{r}\Theta}(r,\ \theta)$ is the joint PDF of r and Θ, then by definition [Eq. (10.44)], the probability of observing r and Θ in this region is $p_{\mathrm{r}\Theta}(r,\ \theta)\,dr d\theta$. But we also know that this probability is $p_{xy}(x,\ y)$ times the area $r\,dr d\theta$ of the shaded region.

Hence, [Eq. (10.52)]

$$\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)2\sigma^2}\, r\, dr d\theta = p_{r\Theta}(r,\ \theta)\, dr d\theta$$

and

$$\begin{aligned} p_{r\Theta}(r,\ \theta) &= \frac{r}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2} \\ &= \frac{r}{2\pi\sigma^2}e^{-r^2/2\sigma^2} \end{aligned} \tag{10.53}$$

and [Eq. (10.47a)]

$$p_r(r) = \int_{-\infty}^{\infty} p_{r\Theta}\,(r,\ \theta)\, d\theta$$

Because Θ exists only in the region (0, 2π),

$$\begin{aligned} p_r(r) &= \int_0^{2\pi} \frac{r}{2\pi\sigma^2}e^{-r^2/2\sigma^2}\, d\theta \\ &= \frac{r}{\sigma^2}e^{-r^2/2\sigma^2} \qquad \mathrm{r} \geq 0 \end{aligned} \tag{10.54a}$$

Note that r is always greater than 0. In a similar way, we find

$$p_\Theta(\theta) = \begin{cases} \dfrac{1}{2\pi} & 0 \leq \Theta < 2\pi \\ 0 & \text{otherwise} \end{cases} \tag{10.54b}$$

RVs r and Θ are independent because $p_{r\Theta}(r,\ \theta) = p_r(r)p_\Theta(\theta)$. The PDF $p_r(r)$ is the **Rayleigh density function**. We shall later show that the envelope of narrow-band gaussian noise has a Rayleigh density. Both $p_r(r)$ and $p_\Theta(\theta)$ are shown in Fig. 10.13b and c.

10.3 STATISTICAL AVERAGES (MEANS)

Averages are extremely important in the study of RVs. In order to find a proper definition for the average of a random variable x, consider the problem of determining the average height of the entire population of a country. Let us assume that we have enough resources to gather data about the height of every person. If the data is recorded within the accuracy of an inch, then the height x of every person will be approximated to one of the n numbers $x_1,\ x_2,\ \ldots,\ x_n$. If there are N_i persons of height x_i, then the average height $\overline{\mathrm{x}}$ is given by

$$\overline{\mathrm{x}} = \frac{N_1x_1 + N_2x_2 + \cdots + N_nx_n}{N}$$

where the total number of persons is $N = \sum_i N_i$. Hence,

$$\overline{\mathrm{x}} = \frac{N_1}{N}x_1 + \frac{N_2}{N}x_2 + \cdots + \frac{N_n}{N}x_n$$

In the limit as $N \to \infty$, the ratio N_i/N approaches $P_x(x_i)$ according to the relative-frequency definition of the probability. Hence,

$$\overline{\mathrm{x}} = \sum_{i=1}^{n} x_i P_x(x_i)$$

The mean value is also called the **average value**, or **expected value**, of the RV x and is denoted by $E[\mathrm{x}]$. Thus,

$$\overline{\mathrm{x}} = E[\mathrm{x}] = \sum_{i} x_i P_x(x_i) \tag{10.55a}$$

We shall use both these notations, depending on the circumstances.

If the RV x is continuous, an argument similar to that used in arriving at Eq. (10.55a) yields

$$\overline{\mathrm{x}} = E[\mathrm{x}] = \int_{-\infty}^{\infty} x p_{\mathrm{x}}(x)\, dx \tag{10.55b}$$

This result can be derived by approximating the continuous variable x with a discrete variable by quantizing it in steps of Δx and then letting $\Delta x \to 0$.

Equation (10.55b) is more general and includes Eq. (10.55a), because the discrete RV can be considered as a continuous RV with an impulsive density. In such a case, Eq. (10.55b) reduces to Eq. (10.55a).

As an example, consider the general gaussian PDF given by (Fig. 10.10)

$$p_{\mathrm{x}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \tag{10.56a}$$

From Eq. (10.55b) we have

$$\overline{\mathrm{x}} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-(x-m)^2/2\sigma^2}\, dx$$

Changing the variable to $x = y + m$ yields

$$\begin{aligned}\overline{\mathrm{x}} &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (y+m) e^{-y^2/2\sigma^2}\, dy \\ &= \frac{1}{\sigma\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} y e^{-y^2/2\sigma^2}\, dy + m \int_{-\infty}^{\infty} e^{-y^2/2\sigma^2}\, dy\right]\end{aligned}$$

The first integral inside the bracket is zero, because the integrand is an odd function of y. The second integral is found from standard tables[3] to be $\sigma\sqrt{2\pi}$. Hence,

$$\overline{\mathrm{x}} = m \tag{10.56b}$$

Mean of a Function of a Random Variable

It is often necessary to find the mean value of a function of a RV. For instance, in practice we are often interested in the mean square amplitude of a signal. The mean square amplitude is the mean of the square of the amplitude x, that is, $\overline{\mathrm{x}^2}$.

In general, we may seek the mean value of an RV y that is a function of the RV x; that is, we wish to find $\overline{\mathrm{y}}$ where $\mathrm{y} = g(\mathrm{x})$. Let x be a discrete RV that takes values $x_1, x_2, \ldots, x_n$ with probabilities $P_{\mathrm{x}}(x_1), P_{\mathrm{x}}(x_2), \ldots, P_{\mathrm{x}}(x_n)$, respectively. But because $y = g(\mathrm{x})$, y takes values $g(x_1),\ g(x_2),\ \ldots,\ g(x_n)$ with probabilities $P_{\mathrm{x}}(x_1),\ P_{\mathrm{x}}(x_2),\ \ldots,\ P_{\mathrm{x}}(x_n)$, respectively. Hence, from Eq. (10.55a) we have

$$\overline{\mathrm{y}} = \overline{g(\mathrm{x})} = \sum_{i=1}^{n} g(x_i) P_{\mathrm{x}}(x_i) \tag{10.57a}$$

If x is a continuous RV, a similar line of reasoning leads to

$$\overline{g(\mathrm{x})} = \int_{-\infty}^{\infty} g(x) p_{\mathrm{x}}(x)\, dx \tag{10.57b}$$

EXAMPLE 10.15 A sinusoid generator output voltage is $A \cos \omega t$. This output is sampled randomly (Fig. 10.14a). The sampled output is an RV x, which can take on any value in the range $(-A,\ A)$. Determine the mean value $(\overline{\mathrm{x}})$ and the mean square value $(\overline{\mathrm{x}^2})$ of the sampled output x.

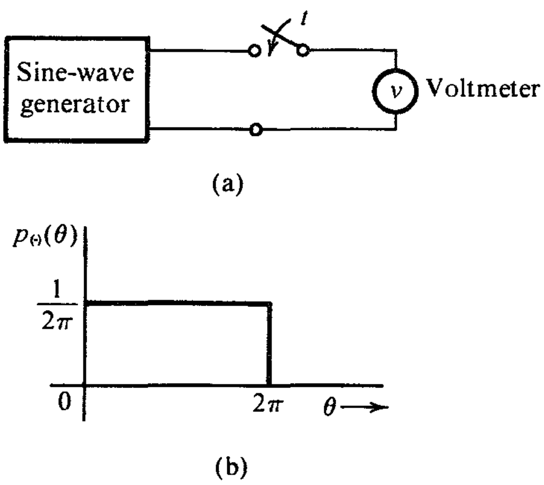


Figure 10.14 Random sampling of a sine-wave generator.

If the output is sampled at a random instant t, the output x is a function of the RV t:

$$\mathrm{x}(t) = A \cos \omega t$$

If we let $\omega t = \Theta$, Θ is also an RV, and if we consider only modulo 2π values of Θ, then the RV Θ lies in the range $(0,\ 2\pi)$. Because t is randomly chosen, Θ can take any value in the range $(0,\ 2\pi)$ with uniform probability. Because the area under the PDF must be unity, $p_\Theta(\theta)$ is as shown in Fig. 10.14b.

The RV x is thus a function of another RV, Θ,

$$\mathrm{x} = A \cos \Theta$$

Hence, from Eq. (10.57b),

$$\overline{\mathrm{x}} = \int_0^{2\pi} x\, p_\Theta(\theta)\, d\theta = \frac{1}{2\pi}\int_0^{2\pi} A \cos\theta\, d\theta = 0$$

and

$$\begin{aligned}\overline{\mathrm{x}^2} &= \int_0^{2\pi} x^2 p_\Theta(\theta)\, d\theta \\ &= \frac{A^2}{2\pi}\int_0^{2\pi} \cos^2\theta\, d\theta = \frac{A^2}{2}\end{aligned}$$

Similarly, for the case of two variables x and y, we have

$$\overline{g(\mathrm{x},\ \mathrm{y})} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,\ y)p_{\mathrm{xy}}(x,\ y)\,dx\,dy \tag{10.58}$$

Mean of the Sum: If $g_1(\mathrm{x},\ \mathrm{y}),\ g_2(\mathrm{x},\ \mathrm{y}),\ \ldots,\ g_n(\mathrm{x},\ \mathrm{y})$ are functions of the RVs x and y, then

$$\overline{g_1(\mathrm{x},\ \mathrm{y}) + g_2(\mathrm{x},\ \mathrm{y}) + \cdots + g_n(\mathrm{x},\ \mathrm{y})} = \overline{g_1(\mathrm{x},\ \mathrm{y})} + \overline{g_2(\mathrm{x},\ \mathrm{y})} + \cdots + \overline{g_n(\mathrm{x},\ \mathrm{y})} \tag{10.59a}$$

The proof is trivial and follows directly from Eq. (10.58).

Thus, the mean (expected value) of the sum is equal to the sum of the means. An important special case is

$$\overline{\mathrm{x}+\mathrm{y}} = \overline{\mathrm{x}} + \overline{\mathrm{y}} \tag{10.59b}$$

Equation (10.59a) can be extended to functions of any number of RVs.

Mean of the Product of Two Functions: Unfortunately, there is no simple result [as Eq. (10.59)] for the product of two functions. For the special case where

$$g(\mathrm{x},\ \mathrm{y}) = g_1(\mathrm{x})g_2(\mathrm{y}) \tag{10.60a}$$

$$\overline{g_1(\mathrm{x})g_2(\mathrm{y})} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(\mathrm{x})g_2(\mathrm{y})p_{\mathrm{xy}}(x,\ y)\,dx\,dy$$

If x and y are independent, then [Eq. (10.50c)]

$$p_{\mathrm{xy}}(x,\ y) = p_{\mathrm{x}}(x)p_{\mathrm{y}}(y)$$

and

$$\overline{g_1(\mathrm{x})g_2(\mathrm{y})} = \int_{-\infty}^{\infty} g_1(x)p_{\mathrm{x}}(x)\,dx \int_{-\infty}^{\infty} g_2(y)p_{\mathrm{y}}(y)\,dy$$

$$= \overline{g_1(\mathrm{x})}\ \overline{g_2(\mathrm{y})} \qquad \text{if x and y independent} \tag{10.60b}$$

A special case of this is

$$\overline{\mathrm{xy}} = \overline{\mathrm{x}}\,\overline{\mathrm{y}} \qquad \text{if x and y independent} \tag{10.60c}$$

Moments

The nth **moment** of an RV x is defined as the mean value of x^n. Thus, the nth moment of x is

$$\overline{\mathrm{x}^n} = \int_{-\infty}^{\infty} x^n p_{\mathrm{x}}(x)\,dx \tag{10.61a}$$

The nth **central moment** of an RV x is defined as

$$\overline{(\mathrm{x}-\overline{\mathrm{x}})^n} = \int_{-\infty}^{\infty} (x-\overline{\mathrm{x}})^n p_{\mathrm{x}}(x)\,dx \tag{10.61b}$$

The second central moment of an RV x is of special importance. It is called the **variance** of x and is denoted by σ_x^2, where σ_x is known as the **standard deviation (S.D.)** of the RV x. By definition,

$$\begin{aligned}\sigma_x^2 &= \overline{(x - \overline{x})^2} \\ &= \overline{x^2} - 2\overline{x\overline{x}} + \overline{x}^2 = \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 \\ &= \overline{x^2} - \overline{x}^2 \qquad (10.62)\end{aligned}$$

Thus, the variance of x is equal to the mean square value minus the square of the mean. When the mean is zero, the variance is the mean square; that is, $\overline{x^2} = \sigma_x^2$.

EXAMPLE 10.16 Find the mean, the mean square, and the variance of the gaussian RV with the PDF in Eq. (10.38) [see Fig. 10.10].

We have

$$\overline{x^2} = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-(x-m)^2/2\sigma^2}\,dx$$

Changing the variable to $y = (x - m)/\sigma$ and integrating, we get

$$\overline{x^2} = \sigma^2 + m^2 \qquad (10.63a)$$

Also, from Eqs. (10.62) and (10.56b),

$$\begin{aligned}\sigma_x^2 &= \overline{x^2} - \overline{x}^2 \\ &= (\sigma^2 + m^2) - (m)^2 \\ &= \sigma^2 \qquad (10.63b)\end{aligned}$$

Hence, a gaussian RV described by the density in Eq. (10.56a) has mean m and variance σ^2. Observe that the gaussian density function is completely specified by the first moment ($\overline{x}$) and the second moment ($\overline{x^2}$).

EXAMPLE 10.17 **Mean Square of the Quantization Error in PCM**

In the PCM scheme discussed in Chapter 6, a signal band-limited to B Hz is sampled at a rate of $2B$ samples per second. The entire range $(-m_p,\ m_p)$ of the signal amplitudes is partitioned into L uniform intervals, each of magnitude $2m_p/L$ (Fig. 10.15a). Each sample is approximated to the midpoint of the interval in which it falls. Thus, sample m in Fig. 10.15a is approximated by a value $\hat{m}$, the midpoint of the interval in which m falls. Each sample is thus approximated (quantized) to one of the L numbers.

The difference $q = m - \hat{m}$ is the quantization error and is an RV. We shall determine $\overline{q^2}$, the mean square value of the quantization error. From Fig. 10.15a it can be seen that q is a continuous RV existing over the range $(-m_p/L,\ m_p/L)$ and is zero outside this range. If we assume that it is equally likely for the sample to lie anywhere in the quantizing interval,* the

---

* Because the quantizing interval is generally very small, variations in the PDF of signal amplitudes over the interval are small and this assumption is reasonable.

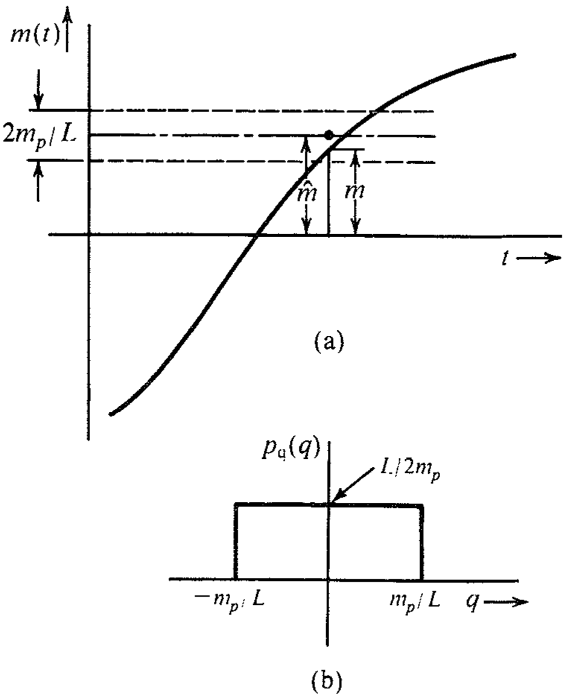


Figure 10.15 Quantization error in PCM and its PDF.

the PDF of q is uniform $[p_q(q) = L/2m_p]$ over the interval $(-m_p/L,\ m_p/L)$, as shown in Fig. 10.15b, and

$$
\begin{aligned}
\overline{q^2} &= \int_{-m_p/L}^{m_p/L} q^2 p_q(q)\, dq \\
&= \frac{L}{2m_p} \frac{q^3}{3} \bigg|_{-m_p/L}^{m_p/L} \\
&= \frac{1}{3}\left(\frac{m_p}{L}\right)^2 \qquad (10.64a)
\end{aligned}
$$

From Fig. 10.15b it can be seen that $\overline{q} = 0$. Hence,

$$
\sigma_q^2 = \overline{q^2} = \frac{1}{3}\left(\frac{m_p}{L}\right)^2 \qquad (10.64b)
$$

EXAMPLE 10.18 Mean Square Error Caused by Channel Noise in PCM

The quantization noise is one of the sources of error in PCM. The other source of error is the channel noise. Each quantized sample is coded by a group of n binary pulses. Because of channel noise, some of these pulses are incorrectly detected at the receiver. Hence, the decoded sample value $\tilde{m}$ at the receiver will differ from the quantized sample value $\hat{m}$ that is transmitted. The error $\varepsilon = \hat{m} - \tilde{m}$ is a random variable. Let us calculate $\overline{\varepsilon^2}$, the mean square error in the sample value caused by the channel noise.

To begin with, let us determine the values that ε can take and the corresponding probabilities. Each sample is transmitted by n binary pulses. The value of ε depends on the position of the incorrectly detected pulse. Consider, for example, the case of $L = 16$ transmitted by four binary pulses ($n = 4$), as shown in Fig. 1.6. Here the transmitted code **1101** represents a value of 13. A detection error in the first digit changes the received code to **0101**, which is a value of 5. This causes an error $\varepsilon = 8$. Similarly, an error in the second digit gives $\varepsilon = 4$. Errors in the third and the fourth digits will give $\varepsilon = 2$ and $\varepsilon = 1$, respectively. In general, the error in the ith digit causes an error $\varepsilon_i = (2^{-i})16$. For a general case, the error $\varepsilon_i = (2^{-i})F$, where F is the full scale, that is, $2m_p$, in PCM. Thus,

$$\varepsilon_i = (2^{-i})(2m_p) \qquad i = 1, 2, \ldots, n$$

Note that the error ε is a discrete RV. Hence,*

$$\overline{\varepsilon^2} = \sum_{i=1}^{n} \varepsilon_i^2 P_\varepsilon(\varepsilon_i) \tag{10.65}$$

Because $P_\varepsilon(\varepsilon_i)$ is the probability that $\varepsilon = \varepsilon_i$, $P_\varepsilon(\varepsilon_i)$ is the probability of error in the detection of the ith digit. Because the error probability of detecting any one digit is the same as that of any other, that is, P_e,

$$\begin{aligned}\overline{\varepsilon^2} &= P_e \sum_{i=1}^{n} \varepsilon_i^2 \\ &= P_e \sum_{i=1}^{n} 4m_p^2(2^{-2i}) \\ &= 4m_p^2 P_e \sum_{i=1}^{n} 2^{-2i}\end{aligned}$$

This summation is a geometric progression with a common ratio $r = 2^{-2}$, with the first term $a_1 = 2^{-2}$ and the last term $a_n = 2^{-2n}$. Hence (see Appendix D.4),

$$\begin{aligned}\overline{\varepsilon^2} &= 4m_p^2 P_e \left[\frac{(2^{-2})2^{-2n} - 2^{-2}}{2^{-2} - 1}\right] \\ &= \frac{4m_p^2 P_e(2^{2n} - 1)}{3(2^{2n})}\end{aligned} \tag{10.66a}$$

Note that the magnitude of the error ε varies from $2^{-1}(2m_p)$ to $2^{-n}(2m_p)$. The error ε can be positive as well as negative. For example, $\varepsilon = 8$ because of a first-digit error in **1101**. But the corresponding error ε will be -8 if the transmitted code is **0101**. Of course the sign of ε does not matter in Eq. (10.65). It must be remembered, however, that ε varies from $-2^{-n}(2m_p)$ to $2^{-2}(2m_p)$ and its probabilities are symmetrical about $\varepsilon = 0$. Hence, $\bar{\varepsilon} = 0$ and

$$\sigma_\varepsilon^2 = \overline{\varepsilon^2} = \frac{4m_p^2 P_e(2^{2n} - 1)}{3(2^{2n})} \tag{10.66b}$$

---

[*] Here we are assuming that the error can occur only in one of the n digits. But more than one digit may be in error. Because the digit error probability $P_e \ll 1$ (on the order 10^{-5} or less), however, the probability of more than one wrong digit is extremely small (see Example 10.6) and its contribution $\varepsilon_i^2 P_\varepsilon(\varepsilon_i)$ is negligible.

Variance of a Sum of Independent Random Variables: The variance of a sum of independent RVs is equal to the sum of their variances. Thus, if x and y are independent RVs and

$$z = x + y$$

then

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 \tag{10.67}$$

This can be shown as follows:

$$\begin{aligned}\sigma_z^2 &= \overline{(z - \bar{z})^2} = \overline{[x + y - (\bar{x} + \bar{y})]^2} \\ &= \overline{[(x - \bar{x}) + (y - \bar{y})]^2} \\ &= \overline{(x - \bar{x})^2} + \overline{(y - \bar{y})^2} + \overline{2(x - \bar{x})(y - \bar{y})} \\ &= \sigma_x^2 + \sigma_y^2 + \overline{2(x - \bar{x})(y - \bar{y})}\end{aligned}$$

Because x and y are independent RVs, $(x - \bar{x})$ and $(y - \bar{y})$ are also independent RVs. Hence, from Eq. (10.60b) we have

$$\overline{(x - \bar{x})(y - \bar{y})} = \overline{(x - \bar{x})}\ \overline{(y - \bar{y})}$$

But

$$\overline{(x - \bar{x})} = \bar{x} - \bar{\bar{x}} = \bar{x} - \bar{x} = 0$$

Similarly,

$$\overline{(y - \bar{y})} = 0$$

and

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

This result can be extended to any number of variables. If RVs x and y both have zero means (i.e., $\bar{x} = \bar{y} = 0$), then $\bar{z} = \bar{x} + \bar{y} = 0$. Also, because the variance equals the mean square value when the mean is zero, it follows that

$$\overline{z^2} = \overline{(x + y)^2} = \overline{x^2} + \overline{y^2} \tag{10.68}$$

provided $\bar{x} = \bar{y} = 0$, and provided x and y are independent RVs.

EXAMPLE 10.19 Total Mean Square Error in PCM

In PCM, as seen in Examples 10.17 and 10.18, a signal sample m is transmitted as a quantized sample $\hat{m}$, causing a quantization error $q = m - \hat{m}$. Because of channel noise, the transmitted sample $\hat{m}$ is read as $\tilde{m}$, causing a detection error $\epsilon = \hat{m} - \tilde{m}$. Hence, the actual signal sample m is received as $\tilde{m}$ with a total error

$$m - \tilde{m} = (m - \hat{m}) + (\hat{m} - \tilde{m}) = q + \epsilon$$

where both q and ϵ are zero mean RVs. Because the quantization error q and the channel-noise error ϵ are independent, the mean square of the sum is [see Eq. (10.68)]

$$\overline{(\mathrm{m}-\tilde{\mathrm{m}})^2} = \overline{(\mathrm{q}+\epsilon)^2} = \overline{\mathrm{q}^2} + \overline{\epsilon^2}$$

$$= \frac{1}{3}\left(\frac{m_p}{L}\right)^2 + \frac{4m_p^2 P_\epsilon(2^{2n}-1)}{3(2^{2n})}$$

Also, because $L = 2^n$,

$$\overline{(\mathrm{m}-\tilde{\mathrm{m}})^2} = \overline{\mathrm{q}^2} + \overline{\epsilon^2} = \frac{{m_p}^2}{3(2^{2n})}[1 + 4P_\epsilon(2^{2n}-1)] \tag{10.69}$$

Chebyshev's Inequality

The standard deviation σ_{x} of an RV x is a measure of the width of its PDF. The larger the σ_{x}, the wider the PDF. Figure 10.16 illustrates this effect for a gaussian PDF. Chebyshev's inequality is a statement of this fact. It states that for a zero mean RV x

$$P(|\mathrm{x}| \le k\sigma_{\mathrm{x}}) \ge 1 - \frac{1}{k^2} \tag{10.70}$$

This means the probability of observing x within a few standard deviations is very high. For example, the probability of finding $|\mathrm{x}|$ within $3\sigma_{\mathrm{x}}$ is equal to or greater than 0.88. Thus, for a PDF with $\sigma_{\mathrm{x}} = 1$, $P(|\mathrm{x}| \le 3) \ge 0.88$, whereas for a PDF with $\sigma_{\mathrm{x}} = 3$, $P(|\mathrm{x}| \le 9) \ge 0.88$. It is clear that the PDF with $\sigma_{\mathrm{x}} = 3$ is spread out much more than the PDF with $\sigma_{\mathrm{x}} = 1$. Hence, σ_{x} or σ_{x}^2 is often used as a measure of the width of a PDF. In Chapter 12, we shall use this measure to estimate the bandwidth of a signal spectrum. The proof of Eq. (10.70) is as follows:

$$\sigma_{\mathrm{x}}^2 = \int_{-\infty}^{\infty} x^2 p_{\mathrm{x}}(x)\,dx$$

Because the integrand is positive,

$$\sigma_{\mathrm{x}}^2 \ge \int_{|x|\ge k\sigma_{\mathrm{x}}} x^2 p_{\mathrm{x}}(x)\,dx$$

If we replace x by its smallest value $k\sigma_{\mathrm{x}}$, the inequality still holds,

$$\sigma_{\mathrm{x}}^2 \ge k^2\sigma_{\mathrm{x}}^2 \int_{|x|\ge k\sigma_{\mathrm{x}}} p_{\mathrm{x}}(x)\,dx = k^2\sigma_{\mathrm{x}}^2 P(|\mathrm{x}| \ge k\sigma_{\mathrm{x}})$$

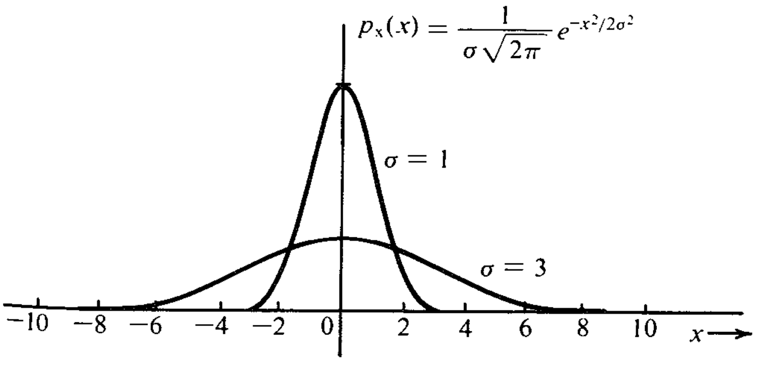


Figure 10.16 Gaussian PDF with standard deviation $\sigma = 1$ and $\sigma = 3$.

or

$$P(|\mathrm{x}| \geq k\sigma_{\mathrm{x}}) \leq \frac{1}{k^2}$$

Hence,

$$P(|\mathrm{x}| \leq k\sigma_{\mathrm{x}}) \geq 1 - \frac{1}{k^2}$$

This inequality can be generalized for a nonzero mean RV as:

$$P(|\mathrm{x} - \bar{\mathrm{x}}| \leq k\sigma_{\mathrm{x}}) \geq 1 - \frac{1}{k^2} \tag{10.71}$$

EXAMPLE 10.20 Estimate the width, or spread, of a gaussian PDF [Eq. (10.56a)].

For a gaussian RV [see Eqs. (10.34) and (10.39b)]

$$\begin{aligned} P(|\mathrm{x} - \bar{\mathrm{x}}| < \sigma) &= 1 - 2Q(1) = 0.6826 \\ P(|\mathrm{x} - \bar{\mathrm{x}}| < 2\sigma) &= 1 - 2Q(2) = 0.9546 \\ P(|\mathrm{x} - \bar{\mathrm{x}}| < 3\sigma) &= 1 - 2Q(3) = 0.9974 \end{aligned}$$

This means that the area under the PDF over the interval $(\bar{\mathrm{x}} - 3\sigma, \bar{\mathrm{x}} + 3\sigma)$ is 99.74% of the total area. A negligible fraction (0.26%) of the area lies outside this interval. Hence, the width, or spread, of the gaussian PDF may be considered roughly $\pm 3\sigma$ about its mean, giving a total width of roughly 6σ.

10.4 CENTRAL-LIMIT THEOREM

Under certain conditions, the sum of a large number of independent RVs tends to be a gaussian random variable, independent of the probability densities of the variables added.* The rigorous statement of this tendency is what is known as the **central-limit theorem**.† Proof of this theorem can be found in the References.[4,5] We shall give here only a simple plausibility argument.

Consider a sum of two RVs x and y:

$$\mathrm{z} = \mathrm{x} + \mathrm{y}$$

Because $\mathrm{z} = \mathrm{x} + \mathrm{y}$, $\mathrm{y} = \mathrm{z} - \mathrm{x}$ regardless of the value of x. Hence, the event $\mathrm{z} \leq z$ is the joint event [$\mathrm{y} \leq z - x$ and x to have any value in the range $(-\infty, \infty)$]. Hence,

$$\begin{aligned} F_{\mathrm{z}}(z) &= P(\mathrm{z} \leq z) = P(\mathrm{x} \leq \infty,\ \mathrm{y} \leq z - x) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} p_{\mathrm{xy}}(x,\ y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} p_{\mathrm{xy}}(x,\ y)\,dy \end{aligned}$$

---

* If the variables are gaussian, this is true even if the variables are not independent.

† Actually, a group of theorems collectively called the central-limit theorem.

and

$$p_z(z) = \frac{dF_z(z)}{dz} = \int_{-\infty}^{\infty} p_{xy}(x,\ z-x)\,dx$$

If x and y are independent RVs, then

$$p_{xy}(x,\ z-x) = p_x(x)p_y(z-x)$$

and

$$p_z(z) = \int_{-\infty}^{\infty} p_x(x)p_y(z-x)\,dx \tag{10.72}$$

The PDF $p_z(z)$ is the convolution of PDFs $p_x(x)$ and $p_y(y)$. We can extend this result to a sum of n independent RVs $x_1,\ x_2,\ \ldots,\ x_n$. If

$$z = x_1 + x_2 + \cdots + x_n$$

then the PDF $p_z(z)$ will be the convolution of PDFs $p_{x_1}(x_1),\ p_{x_2}(x_2),\ \ldots,\ p_{x_n}(x_n)$.

The tendency toward a gaussian distribution when a large number of functions are convolved in shown in Fig. 10.17. For simplicity, we assume all PDFs to be identical, that is, a gate function 0.5 rect $(x/2)$. Figure 10.17 shows the successive convolutions of gate functions. The tendency toward a bell-shaped density is evident.

10.5 CORRELATION

Often we are interested in determining the nature of dependence between two entities, such as smoking and lung cancer. Consider a random experiment with two outcomes described by RVs x and y. We conduct several trials of this experiment and record values of x and y for each trial. From this data, it may be possible to determine the nature of a dependence between x and y. The covariance of RVs x and y is one measure that is simple to compute and can yield useful information about the dependence between x and y.

The covariance σ_{xy} of two RVs is defined as

$$\sigma_{xy} = \overline{(x-\bar{x})(y-\bar{y})} \tag{10.73}$$

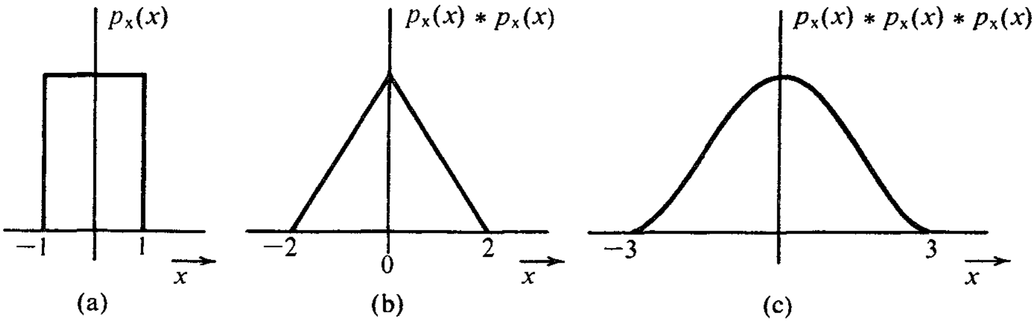


Figure 10.17 Demonstration of the central-limit theorem.

Note that the concept of covariance is a natural extension of the concept of variance, which is defined as

$$\sigma_{\mathrm{x}}^2 = \overline{(\mathrm{x} - \bar{\mathrm{x}})(\mathrm{x} - \bar{\mathrm{x}})}$$

Let us consider a case where the variables x and y are dependent such that they tend to vary in harmony; that is, if x increases y increases, and if x decreases y also decreases. For instance, x may be the average daily temperature of a city and y the volume of soft drink sales that day in the city. It is reasonable to expect the two quantities to vary in harmony for a majority of the cases. Suppose we consider the following random experiment: pick a random day and record the average temperature of that day as the value of x and the soft drink sales volume that day as the value of y. We perform this measurement over several days (several trials of the random experiment) and record the data x and y for each trial. We now plot points $(x,\ y)$ for all the trials. This plot, known as the **scatter diagram**, may appear as shown in Fig. 10.18a. The plot shows that when x is large, y is likely to be large. Note the use of the word *likely*. It is not *always* true that y will be large if x is large, but it is true most of the time. In other words, a few cases will occur when a low average temperature may produce higher soft drink sales due to some atypical situation, such as a big convention. This is quite obvious from the scatter diagram in Fig. 10.18a.

To continue this example, the variable $\mathrm{x} - \bar{\mathrm{x}}$ represents the difference between actual and average values of x, and $\mathrm{y} - \bar{\mathrm{y}}$ represents the difference between actual and average values of y. It is more instructive to plot $(\mathrm{y} - \bar{\mathrm{y}})$ vs. $(\mathrm{x} - \bar{\mathrm{x}})$. This is the same as the scatter diagram in Fig. 10.18a with the origin shifted to $(\bar{\mathrm{x}},\ \bar{\mathrm{y}})$ (see Fig. 10.18b). From this figure, we see that a day with an above-average temperature is likely to produce above average soft drink sales, and a day with a below-average temperature is likely to produce below average soft drink sales. That is, if $x - \bar{\mathrm{x}}$ is positive, $y - \bar{\mathrm{y}}$ is likely to be positive, and if $x - \bar{\mathrm{x}}$ is negative, $y - \bar{\mathrm{y}}$ is more likely to be negative. Thus, the quantity $(x - \bar{\mathrm{x}})(y - \bar{\mathrm{y}})$ will be positive for most trials. We compute this product for every pair, add these products, and then divide by the number of trials.

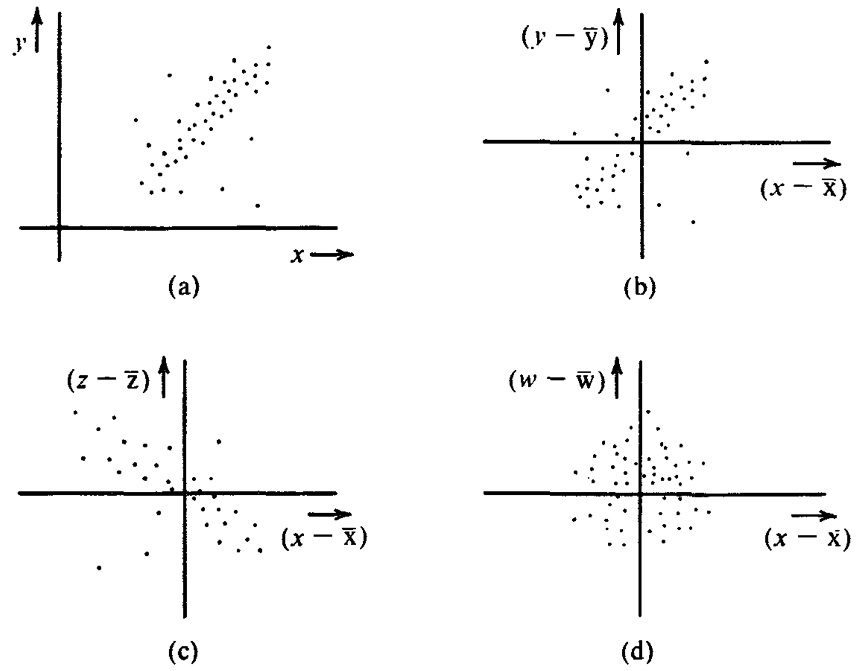


Figure 10.18 Scatter diagrams. (a), (b) Positive correlation. (c) Negative correlation. (d) Zero correlation.

The result is the mean value of $(x-\bar{x})(y-\bar{y})$, that is, the covariance $\sigma_{xy} = \overline{(x-\bar{x})(y-\bar{y})}$. The covariance will be positive in the example under consideration. In such cases, we say that a **positive correlation** exists between variables x and y. We can conclude that a positive correlation implies variation of two variables in harmony (in the same direction, up or down).

Next, we consider the case where the two variables are x, the average daily temperature, and z, the sales volume of sweaters that day. It is reasonable to believe that as x (daily average temperature) increases, z (the sweater sales volume) tends to decrease. A hypothetical scatter diagram for this experiment is shown in Fig. 10.18c. Thus, if $x-\bar{x}$ is positive (above-average temperature), $z-\bar{z}$ is likely to be negative (below-average sweater sales). Similarly, when $x-\bar{x}$ is negative, $z-\bar{z}$ is likely to be positive. The product $(x-\bar{x})(z-\bar{z})$ will be negative for most of the trials, and the mean $\overline{(x-\bar{x})(z-\bar{z})} = \sigma_{xz}$ will be negative. In such a case, we say that **negative correlation** exists between x and y. It should be stressed here that negative correlation does not mean that x and y are unrelated. It means that they are dependent, but when one increases, the other decreases, and vice versa.

Lastly, consider the variables x (the average daily temperature) and w (the number of child births). It is reasonable to expect that the daily temperature has little to do with the number of children born. A hypothetical scatter diagram for this case will appear as shown in Fig. 10.18d. If $x-\bar{x}$ is positive, $w-\bar{w}$ is equally likely to be positive or negative. The product $(x-\bar{x})(w-\bar{w})$ is therefore equally likely to be positive or negative, and the mean $\overline{(x-\bar{x})(w-\bar{w})} = \sigma_{xw}$ will be zero. In such a case, we say that RVs x and w are **uncorrelated.**

To reiterate, if σ_{xy} is positive (or negative), then x and y are said to have a positive (or negative) correlation, and if $\sigma_{xy} = 0$, then the variables x and y are said to be uncorrelated.

From this discussion, it appears that under suitable conditions, covariance can serve as a measure of the dependence of two variables. It often provides *some* information about the interdependence of the two RVs and proves useful in a number of applications.

The covariance σ_{xy} may be expressed in another way, as follows. By definition,

$$\begin{aligned}\sigma_{xy} &= \overline{(x-\bar{x})(y-\bar{y})} \\ &= \overline{xy} - \overline{\bar{x}y} - \overline{x\bar{y}} + \overline{\bar{x}\bar{y}} \\ &= \overline{xy} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\ &= \overline{xy} - \bar{x}\bar{y} \end{aligned} \tag{10.74}$$

From Eq. (10.74) it follows that the variables x and y are uncorrelated ($\sigma_{xy} = 0$) if

$$\overline{xy} = \bar{x}\bar{y} \tag{10.75}$$

Note that for independent RVs [Eq. (10.60c)]

$$\overline{xy} = \bar{x}\bar{y} \qquad \text{and} \qquad \sigma_{xy} = 0$$

Hence, independent RVs are uncorrelated. This supports the heuristic argument presented earlier. It should be noted that whereas independent variables are uncorrelated, the converse is not necessarily true—uncorrelated variables are not necessarily independent (see Prob. 10.5-3). Independence is, in general, a stronger and more restrictive condition than uncorrelatedness. For independent variables, we have shown [Eq. (10.60b)] that

$$\overline{g_1(x)g_2(y)} = \overline{g_1(x)}\;\overline{g_2(y)}$$

for any functions g_1 and g_2, whereas for uncorrelatedness, the only requirement is that

$$\overline{\mathrm{xy}} = \bar{\mathrm{x}}\,\bar{\mathrm{y}}$$

The **coefficient of correlation** ρ_{xy} is σ_{xy} normalized by $\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}$,

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} \tag{10.76}$$

It can be shown that (see Prob. 10.5-1) that

$$-1 \le \rho_{\mathrm{xy}} \le 1 \tag{10.77}$$

Mean Square of the Sum of Uncorrelated Variables

If x and y are uncorrelated, then for $\mathrm{z} = \mathrm{x} + \mathrm{y}$ we show that

$$\sigma_{\mathrm{z}}^2 = \sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2 \tag{10.78}$$

That is, the variance of the sum is the sum of variances for uncorrelated RVs. We have proved this result earlier for independent variables x and y. Following the development after Eq. (10.67), we have

$$\begin{aligned}\sigma_{\mathrm{z}}^2 &= \overline{[(\mathrm{x}-\bar{\mathrm{x}}) + (\mathrm{y}-\bar{\mathrm{y}})]^2} \\ &= \overline{(\mathrm{x}-\bar{\mathrm{x}})^2} + \overline{(\mathrm{y}-\bar{\mathrm{y}})^2} + 2\overline{(\mathrm{x}-\bar{\mathrm{x}})(\mathrm{y}-\bar{\mathrm{y}})} \\ &= \sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2 + 2\sigma_{\mathrm{xy}}\end{aligned}$$

Because x and y are uncorrelated, $\sigma_{\mathrm{xy}} = 0$, and Eq. (10.78) follows. If x and y have zero means, then z also has a zero mean, and the mean square values of these variables are equal to their variances. Hence,

$$\overline{(\mathrm{x}+\mathrm{y})^2} = \overline{\mathrm{x}^2} + \overline{\mathrm{y}^2} \tag{10.79}$$

if x and y are uncorrelated and have zero means. Thus, Eqs. (10.78) and (10.79) are valid not only when x and y are independent, but also under the less restrictive condition that x and y be uncorrelated.

10.6 LINEAR MEAN SQUARE ESTIMATION

When two random variables x and y are related (or dependent), then a knowledge of one gives certain information about the other. Hence, it is possible to estimate the value of y from a knowledge of the value of x. The estimate of y will be another random variable $\hat{\mathrm{y}}$. The estimated value $\hat{y}$ will in general be different from the actual value y. One may choose various criteria of goodness for estimation. Minimum mean square error is one possible criterion. The optimum estimate in this case minimizes the mean square error $\overline{\epsilon^2}$ given by

$$\overline{\epsilon^2} = \overline{(\mathrm{y}-\hat{\mathrm{y}})^2}$$

In general, the optimum estimate $\hat{\mathrm{y}}$ is a nonlinear function of x.* We simplify the problem by constraining the estimate $\hat{\mathrm{y}}$ to be a linear function of x of the form

$$\hat{\mathrm{y}} = a\mathrm{x}$$

assuming that $\overline{\mathrm{x}} = 0$.† In this case,

$$\begin{aligned}\overline{\epsilon^2} &= \overline{(\mathrm{y} - \hat{\mathrm{y}})^2} = \overline{(\mathrm{y} - a\mathrm{x})^2} \\ &= \overline{\mathrm{y}^2} + a^2\overline{\mathrm{x}^2} - 2a\overline{\mathrm{xy}}\end{aligned}$$

To minimize $\overline{\epsilon^2}$, we have

$$\frac{\partial\overline{\epsilon^2}}{\partial a} = 2a\overline{\mathrm{x}^2} - 2\overline{\mathrm{xy}} = 0$$

Hence,

$$a = \frac{\overline{\mathrm{xy}}}{\overline{\mathrm{x}^2}} = \frac{R_{\mathrm{xy}}}{R_{\mathrm{xx}}} \tag{10.80}$$

Where $R_{xy} = \overline{\mathrm{xy}}$, $R_{xx} = \overline{x^2}$, and $R_{yy} = \overline{y^2}$. Note that for this value of a,

$$\epsilon = \mathrm{y} - a\mathrm{x} = \mathrm{y} - \frac{R_{\mathrm{xy}}}{R_{\mathrm{xx}}}\mathrm{x}$$

Hence,

$$\begin{aligned}\overline{\mathrm{x}\epsilon} &= \overline{\mathrm{x}\left(\mathrm{y} - \frac{R_{\mathrm{xy}}}{R_{\mathrm{xx}}}\mathrm{x}\right)} \\ &= \overline{\mathrm{xy}} - \frac{R_{\mathrm{xy}}}{R_{\mathrm{xx}}}\overline{\mathrm{x}^2}\end{aligned}$$

Since by definition $\overline{\mathrm{xy}} = R_{\mathrm{xy}}$ and $\overline{\mathrm{xx}} = \overline{\mathrm{x}^2} = R_{\mathrm{xx}}$, we have

$$\overline{\mathrm{x}\epsilon} = R_{\mathrm{xy}} - R_{\mathrm{xy}} = 0 \tag{10.81}$$

Hence, the data (x) and the error (ϵ) are orthogonal (implying uncorrelatedness in this case) when the mean square error is minimum.

The mean-square error is given by

$$\begin{aligned}\overline{\epsilon^2} &= \overline{(\mathrm{y} - a\mathrm{x})^2} \\ &= \overline{\mathrm{y}^2} - 2a\overline{\mathrm{xy}} + a^2\overline{\mathrm{x}^2} \\ &= R_{\mathrm{yy}} - \frac{2R_{\mathrm{xy}}^2}{R_{\mathrm{xx}}} + \frac{R_{\mathrm{xy}}^2}{R_{\mathrm{xx}}} \\ &= R_{\mathrm{yy}} - \frac{R_{\mathrm{xy}}^2}{R_{\mathrm{xx}}} = R_{\mathrm{yy}} - aR_{\mathrm{xy}}\end{aligned} \tag{10.82}$$

---

* It can be shown that[5] the optimum estimate $\hat{\mathrm{y}}$ is the conditional mean of y when $\mathrm{x} = x$, that is,

$$\hat{\mathrm{y}} = E[\mathrm{y} \mid \mathrm{x} = x]$$

In general, this is a nonlinear function of x.

† Throughout the discussion, the variables x, y, ... will be assumed to have zero mean. This can be done without loss of generality. If the variables have nonzero means, we can form new variables $\mathrm{x}' = \mathrm{x} - \overline{\mathrm{x}}$ and $\mathrm{y}' = \mathrm{y} - \overline{\mathrm{y}}$, and so on. The new variables obviously have zero mean values.

Estimation of a Random Variable Using n Random Variables: If a random variable x_0 is related to n RVs $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$, then we can estimate x_0 using a linear combination* of $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$:

$$\begin{aligned}\hat{\mathrm{x}}_0 &= a_1\mathrm{x}_1 + a_2\mathrm{x}_2 + \cdots + a_n\mathrm{x}_n \\ &= \sum_{i=1}^{n} a_i\mathrm{x}_i \end{aligned} \tag{10.83}$$

The mean square error is given by

$$\overline{\epsilon^2} = \overline{[\mathrm{x}_0 - (a_1\mathrm{x}_1 + a_2\mathrm{x}_2 + \cdots + a_n\mathrm{x}_n)]^2}$$

To minimize $\overline{\epsilon^2}$, we must set

$$\frac{\partial\overline{\epsilon^2}}{\partial a_1} = \frac{\partial\overline{\epsilon^2}}{\partial a_2} = \cdots = \frac{\partial\overline{\epsilon^2}}{\partial a_n} = 0$$

that is,

$$\frac{\partial\overline{\epsilon^2}}{\partial a_i} = \frac{\partial}{\partial a_i}\overline{[\mathrm{x}_0 - (a_1\mathrm{x}_1 + a_2\mathrm{x}_2 + \cdots + a_n\mathrm{x}_n)]^2} = 0$$

Interchanging the order of differentiation and averaging, we have

$$\frac{\partial\overline{\epsilon^2}}{\partial a_i} = -2\overline{[\mathrm{x}_0 - (a_1\mathrm{x}_1 + a_2\mathrm{x}_2 + \cdots + a_n\mathrm{x}_n)]\mathrm{x}_i} = 0 \tag{10.84}$$

or

$$R_{0i} = a_1R_{i1} + a_2R_{i2} + \cdots + a_nR_{in} \tag{10.85}$$

where

$$R_{ij} = \overline{\mathrm{x}_i\mathrm{x}_j}$$

differentiating $\overline{\epsilon^2}$ with respect to $a_1, a_2, \ldots, a_n$ and equating to zero, we obtain n simultaneous equations of the form shown in Eq. (10.85). The desired constants $a_1, a_2, \ldots, a_n$ can be found from these equations by using Cramer's rule as

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ R_{n1} & R_{n2} & \cdots & R_{nn} \end{bmatrix}^{-1} \begin{bmatrix} R_{01} \\ R_{02} \\ \vdots \\ R_{0n} \end{bmatrix} \tag{10.86}$$

These are the Yule-Walker equations. Equation (10.84) shows that $\overline{\epsilon\mathrm{x}_i} = 0$ ($i = 1, 2, \ldots, n$). This means ϵ (the error) is orthogonal to data ($\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$) for optimum estimation. The mean square error (under optimum conditions) is

$$\overline{\epsilon^2} = \overline{\epsilon\epsilon} = \overline{\epsilon[\mathrm{x}_0 - (a_1\mathrm{x}_1 + a_2\mathrm{x}_2 + \cdots + a_n\mathrm{x}_n)]}$$

---

* Throughout this section as before, we assume that all the random variables have zero mean values. This can be done without loss of generality.

Because $\overline{\epsilon x_i} = 0 \ (i = 1, 2, \ldots, n)$,

$$\begin{aligned}\overline{\epsilon^2} &= \overline{\epsilon x_0} \\ &= \overline{x_0[x_0 - (a_1 x_1 + a_2 x_2 + \cdots + a_n x_n)]} \\ &= R_{00} - (a_1 R_{01} + a_2 R_{02} + \cdots + a_n R_{0n}) \qquad (10.87)\end{aligned}$$

EXAMPLE 10.21 In differential pulse code modulation (DPCM), instead of transmitting sample values directly, we estimate (predict) the value of each sample from the knowledge of previous n samples. The estimation error ϵ_k, the difference between the actual value and the estimated value of the kth sample, is quantized and transmitted (Fig. 10.19). Because the estimation error ϵ_k is smaller than the sample value m_k, for the same number of quantization levels (the same number of PCM code bits), the SNR is increased. It was shown in Chapter 6 that the SNR improvement is equal to $\overline{m^2}/\overline{\epsilon^2}$, where $\overline{m^2}$ and $\overline{\epsilon^2}$ are the mean square values of the speech signal and the estimation error ϵ, respectively. In this example, we shall find the optimum linear second-order predictor and the corresponding SNR improvement.

The equation of a second-order estimator (predictor), shown in Fig. 10.19, is

$$\hat{m}_k = a_1 m_{k-1} + a_2 m_{k-2}$$

where $\hat{m}_k$ is the best linear estimate of m_k. The estimation error ϵ_k is given by

$$\epsilon_k = \hat{m}_k - m_k = a_1 m_{k-1} + a_2 m_{k-2} - m_k$$

For speech signals, Jayant and Noll[6] give the values of correlations of various samples as: $\overline{m_k m_k} = \overline{m^2}$, $\overline{m_k m_{k-1}} = 0.825\overline{m^2}$, $\overline{m_k m_{k-2}} = 0.562\overline{m^2}$, $\overline{m_k m_{k-3}} = 0.308\overline{m^2}$, $\overline{m_k m_{k-4}} = 0.004\overline{m^2}$, and $\overline{m_k m_{k-5}} = -0.243\overline{m^2}$.

Note that $R_{ij} = \overline{m_k m_{k-(j-i)}}$. Hence, $R_{11} = R_{22} = \overline{m_2}$, $R_{12} = R_{21} = R_{01} = 0.825\overline{m^2}$, and $R_{02} = 0.562\overline{m^2}$.

The optimum values of a_1 and a_2 are found from Eq. (10.86) as $a_1 = 1.1314$ and $a_2 = -0.3714$, and the mean square error in the estimation is given by Eq. (10.87) as

$$\overline{\epsilon^2} = [1 - (0.825a_1 + 0.562a_2)]\overline{m^2} = 0.2753\overline{m^2}$$

The SNR improvement is $10 \log_{10} \overline{m^2}/0.2752\overline{m^2} = 5.6$ dB.

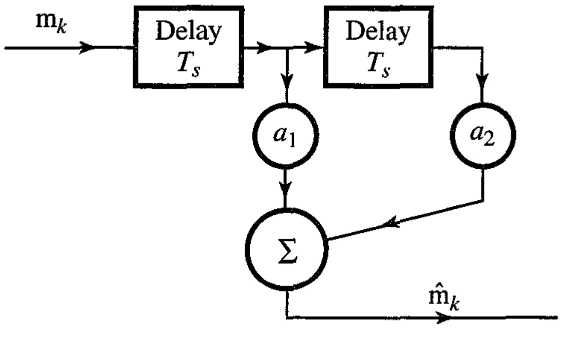


Figure 10.19 Second-order predictor in Example 10.21.

REFERENCES

1. J. Singh, *Great Ideas of Modern Mathematics*, Dover, Boston, MA, 1959.
2. M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964, sec. 26.
3. The Chemical Rubber Co., *CRC Standard Mathematical Tables*, 26th ed., 1980.
4. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965, p. 83.
5. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, 1995
6. N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

PROBLEMS

10.1-1 A card is drawn randomly from a regular deck of cards. Assign probability to the event that the card drawn is: **(a)** a red card; **(b)** a black queen; **(c)** a picture card (count an ace as a picture card); **(d)** a number card with number 7; **(e)** a number card with number ≤ 5.

10.1-2 Three regular dice are thrown. Assign probabilities to the following events: the sum of the points appearing on the three dice is: **(a)** 4; **(b)** 9; **(c)** 15.

10.1-3 The probability that the number i appears on a throw of a certain loaded dice is k_i $(i = 1, 2, \ldots, 6)$. Assign probabilities to all six outcomes.

10.1-4 A bin contains three oscillator microchips, marked O_1 and O_2, and O_3, and two PLL microchips, marked P_1 and P_2. Two chips are picked randomly in succession without replacement.

(a) How many outcomes are possible, that is, how many points are in the sample space? List all the outcomes and assign probabilities to each of them.

(b) Express the following events as unions of the outcomes in part **(a)** **(i)** one chip drawn is oscillator and the other is PLL; **(ii)** both chips are PLL; **(iii)** both chips are oscillator; and **(iv)** both chips are of the same kind. Assign probabilities to each of these events.

10.1-5 Find the probabilities in Prob. 10.1-4, part **(b)**, using Eq. (10.12).

10.1-6 In Prob. 10.1-4, determine the probability that:

(a) The second pick is an oscillator chip given that the first pick is a PLL chip.

(b) The second pick is an oscillator chip given that the first pick is also an oscillator chip.

10.1-7 A binary source generates digits **1** and **0** randomly with equal probability. Assign probabilities to the following events. In ten digits generated by the source **(a)** there are exactly two **1**'s and eight **0**'s; **(b)** there are at least four **0**'s.

10.1-8 In the California lottery (Lotto), a player chooses any 6 numbers out of 49 numbers (1 through 49). Six balls are drawn randomly (without replacement) from the 49 balls numbered 1 through 49.

(a) Find the probability of matching all 6 balls to the 6 numbers chosen by the player.

(b) Find the probability of matching exactly 5 balls.

(c) Find the probability of matching exactly 4 balls.

(d) Find the probability of matching exactly 3 balls.

10.1-9 A system consists of ten subsystems $s_1, s_2, \ldots, s_{10}$ in cascade (Fig. P10.1-9). If any one of the subsystems fails, the entire system fails. The probability of failure of any one of the subsystems is 0.01.

(a) What is the probability of failure of the system? *Hint:* Consider the probability that none of the subsystems fails.

(b) The reliability of a system is the probability of not failing. If the system reliability is required to be 0.99, what must be the failure probability of each subsystem?

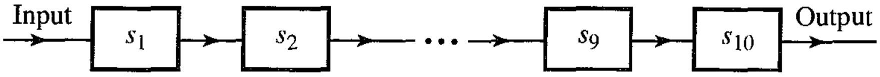


Figure P10.1-9

10.1-10 System reliability improves by using redundant systems. The reliability of the system in Prob. 10.1-9 (Fig. P10.1-9) can be improved by using two such systems in parallel (Fig. P10.1-10). Thus, if one system fails, the other one will still function.

(a) Using the data in Prob. 10.1-9, determine the reliability of the system in Fig. P10.1-10.

(b) If the reliability of this system is required to be 0.999, what must be the failure probability of each subsystem?

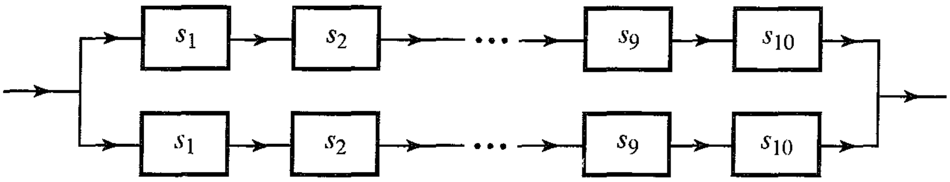


Figure P10.1-10

10.1-11 Compare the reliability of the two systems in Fig. P10.1-11, given that the probability of failure of subsystems s_1 and s_2 is p each.

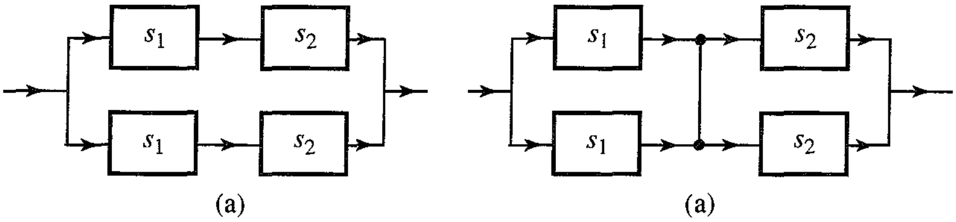


Figure P10.1-11

10.1-12 In a poker game each player is dealt five cards from a regular deck. What is the probability that a player will get a flush (all five cards of one suit)?

10.1-13 Two dice are tossed. One die is regular and the other is biased with probabilities:

$$P(1) = P(6) = \frac{1}{6}, \qquad P(2) = P(4) = 0, \qquad P(3) = P(5) = \frac{1}{3}$$

Determine the probabilities of obtaining a sum: **(a)** 4; **(b)** 5.

10.1-14 In Sec. 10.1, Example 10.5, determine:

(a) $P(B)$, the probability of drawing a red ace in the second draw.

(b) $P(A|B)$, the probability that the first draw was a red ace given that the second draw is a red ace. *Hint*: Event B can occur in two ways: the first draw ia a red ace and the second draw is a red ace, or the first draw is not a red ace and the second draw is a red ace. This is $AB \cup A^cB$ (see Fig. 10.2).

10.1-15 A binary source generates digits **1** and **0** randomly with probabilities $P(\mathbf{1}) = 0.8$ and $P(\mathbf{0}) = 0.2$.

(a) What is the probability that two **1**'s and three **0**'s will occur in a five-digit sequence?

(b) What is the probability that at least three **1**'s will occur in a five-digit sequence?

10.1-16 In a binary communication channel, the receiver detects binary pulses with an error probability P_e. What is the probability that out of 100 received digits, no more than three digits are in error?

10.1-17 A PCM channel is made up of 10 links, with a regenerative repeater at the end of each link. If the detection error probabilities of the 10 detectors are $p_1, p_2, \ldots, p_{10}$, determine the dectection error probability of the entire channel if $p_i \ll 1$.

10.1-18 Example 10.8 in Sec. 10.1 considers the possibility of improving reliability by three repetitions of a digit. Repeat this analysis for five repetitions.

10.1-19 A bin contains nine bad microchips. A good microchip is thrown into the bin by mistake. Someone is trying to retrieve the good chip. He draws a chip randomly and tests it. If the chip is bad he throws it out and draws another chip randomly, repeating the procedure until he finds the good chip.

(a) What is the probability that he will find the good chip in the first trial?

(b) What is the probability that he will find the good chip in five trials?

10.1-20 One out of a group of 10 people is to be selected for a suicide mission by drawing straws. There are 10 straws—nine are of the same length and the tenth is shorter than the others. The straws are drawn one by one by all ten people. The person who draws the shortest straw is selected for the mission. Determine which position in the sequence favors the most and which favors the least drawing the short straw.

10.1-21 A binary source generates digits **1** and **0** randomly with probabilities $P(1) = 0.7$ and $P(0) = 0.3$.

(a) What is the probability that no **1**'s will be generated in a sequence of 10 digits?

(b) What is the probability that eight **1**'s and two **0**'s will be generated in a sequence of 10 digits?

(c) What is the probability that at least five **0**'s will be generated in a sequence of 10 digits?

10.2-1 For a certain binary nonsymmetric channel it is given that

$$P_{\mathrm{y|x}}(\mathbf{0}|\mathbf{1}) = 0.1 \qquad \text{and} \qquad P_{\mathrm{y|x}}(\mathbf{1}|\mathbf{0}) = 0.2$$

where x is the transmitted digit and y is the received digit. If $P_{\mathrm{x}}(\mathbf{0}) = 0.4$, determine $P_{\mathrm{y}}(\mathbf{0})$ and $P_{\mathrm{y}}(\mathbf{1})$.

10.2-2 A binary symmetric channel (see Example 10.6) has an error probability P_e. The probability of transmitting **1** is Q, and the probability of transmitting **0** is $1 - Q$. If the receiver detects an incoming digit as **1**, what is the probability that the corresponding transmitted digit was:

(a) 1; (b) 0? *Hint*: If x is the transmitted digit and y is the received digit, you are given $P_{y|x}(\mathbf{0}|\mathbf{1}) = P_{y|x}(\mathbf{1}|\mathbf{0}) = P_e$. Now using Bayes' rule, find $P_{x|y}(\mathbf{1}|\mathbf{1})$ and $P_{x|y}(\mathbf{0}|\mathbf{1})$.

10.2-3 The PDF of amplitude x of a certain signal $x(t)$ is given by $p_x(x) = 0.5|x|e^{-|x|}$.

(a) Find the probability that $x \geq 1$.

(b) Find the probability that $-1 < x \leq 2$.

(c) Find the probability that $x \leq -2$.

10.2-4 The PDF of an amplitude x of a gaussian signal $x(t)$ is given by

$$p_x(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}$$

This signal is applied to the input of a half-wave rectifier circuit (Fig. P10.2-4). Assuming an ideal diode, determine $F_y(y)$ and $p_y(y)$ of the output signal amplitude y.

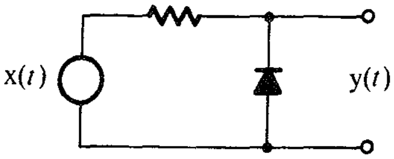


Figure P10.2-4

10.2-5 The PDF of a gaussian variable x is given by

$$p_x(x) = \frac{1}{3\sqrt{2\pi}}e^{-(x-4)^2/18}$$

Determine: **(a)** $P(x \geq 4)$; **(b)** $P(x \geq 0)$; **(c)** $P(x \geq -2)$.

10.2-6 For an RV x with PDF

$$p_x(x) = \frac{1}{2\sqrt{2\pi}}e^{-x^2/32}u(x)$$

(a) Sketch $p_x(x)$ and state (with reasons) if this is a gaussian RV.

(b) Determine: **(i)** $P(x \geq 1)$, **(ii)** $P(1 < x \leq 2)$.

(c) Can you generate RV x from another gaussian RV? Explain.

10.2-7 The joint PDF of RVs x and y is shown in Fig. P10.2-7.

(a) Determine: **(i)** A; **(ii)** $p_x(x)$; **(iii)** $p_y(y)$; **(iv)** $P_{x|y}(x|y)$; **(v)** $P_{y|x}(y|x)$.

(b) Are x and y independent? Explain.

10.2-8 The joint PDF $p_{xy}(x, y)$ of two continuous RVs is given by

$$p_{xy}(x, y) = xy\,e^{-(x^2+y^2)/2}u(x)u(y)$$

(a) Find $p_x(x)$, $p_y(y)$, $p_{x|y}(x|y)$, and $p_{y|x}(y|x)$.

(b) Are x and y independent?

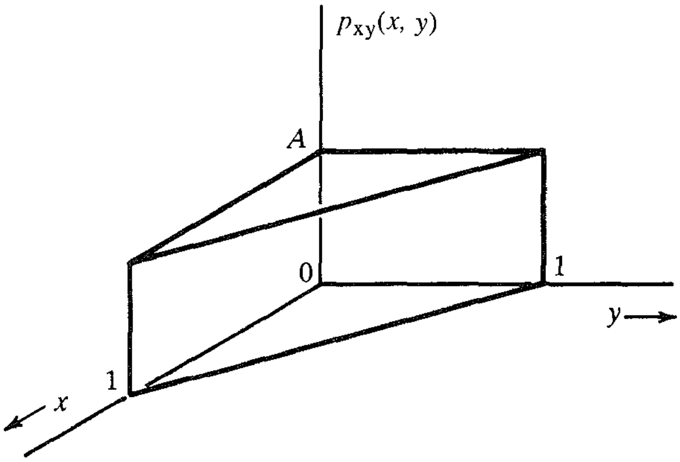


Figure P10.2-7

10.2-9 RVs x and y are said to be jointly gaussian if their joint PDF is given by

$$p_{xy}(x, y) = \frac{1}{2\pi\sqrt{M}} e^{-(ax^2+by^2-2cxy)/2M}$$

where $M = ab - c^2$. Show that $p_x(x)$, $p_y(y)$, $P_{x|y}(x|y)$, and $P_{y|x}(y|x)$ are all gaussian and that $\overline{x^2} = b$, $\overline{y^2} = a$, and $\overline{xy} = c$. *Hint:* Use

$$\int_{-\infty}^{\infty} e^{-px^2+qx}\, dx = \sqrt{\frac{\pi}{p}}\, e^{q^2/4p}$$

10.2-10 The joint PDF of RVs x and y is given by

$$p_{xy}(x, y) = ke^{-(x^2+xy+y^2)}$$

Determine: **(a)** the constant k; **(b)** $p_x(x)$; **(c)** $p_y(y)$; **(d)** $p_{x|y}(x, y)$; **(e)** $p_{y|x}(y|x)$. Are x and y independent?

10.2-11 In the example on threshold detection (Example 10.13), it was assumed that the digits **1** and **0** were transmitted with equal probability. If $P_x(\mathbf{1})$ and $P_x(\mathbf{0})$, the probabilities of transmitting **1** and **0**, respectively, are not equal, show that the optimum threshold is not 0 but is a, where

$$a = \frac{\sigma_n^2}{2A_p} \ln \frac{P_x(\mathbf{0})}{P_x(\mathbf{1})}$$

Hint: Assume that the optimum threshold is a, and write P_e in terms of the Q functions. For the optimum case, $dP_e/da = 0$. Use the fact that

$$Q(x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy$$

and

$$\frac{dQ(x)}{dx} = -\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

10.3-1 If an amplitude x of a gaussian signal x(t) has a mean value of 2 and an rms value of 3, determine its PDF.

10.3-2 Determine the mean, the mean square, and the variance of the RV x in Prob. 10.2-3.

10.3-3 Determine the mean and the mean square value of y in Prob. 10.2-4.

10.3-4 Determine the mean and the mean square value of x in Prob. 10.2-6.

10.3-5 Find the mean, the mean square, and the variance of the RV x whose PDF is shown in Fig. P10.3-5.

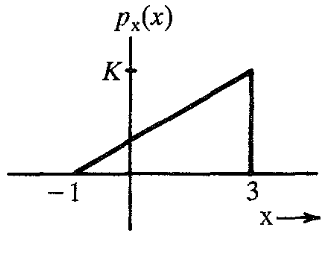


Figure P10.3-5

10.3-6. The sum of the points on two tossed dice is a discrete RV x, as analyzed in Example 10.9. Determine the mean, the mean square, and the variance of x.

10.3-7. For a gaussian PDF $p_x(x) = (1/\sigma_x\sqrt{2\pi})e^{-x^2/2\sigma_x^2}$, show that

$$\overline{x^n} = \begin{cases} (1)(3)(5)\cdots(n-1)\sigma_x^n & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$$

Hint: See appropriate definite integrals in any standard mathematical table.

10.3-8 Ten regular dice are thrown. The sum of the numbers appearing on these ten dice is an RV x. Find $\bar{x}$, $\overline{x^2}$, and σ_x^2. *Hint:* Remember that the outcome of each die is independent.

10.4-1 The random binary signal $x(t)$, shown in Fig. P10.4-1a, can take on only two values, 3 and 0, with equal probability. A gaussian channel noise $n(t)$ shown in Fig. P10.4-1b is added to this signal, giving the received signal $y(t)$. The PDF of the noise amplitude n is gaussian with a zero mean and an rms value of 2. Determine and sketch the PDF of the amplitude y. *Hint*: Use of Eq. (10.72) yields $p_y(y) = p_x(x) * p_n(n)$.

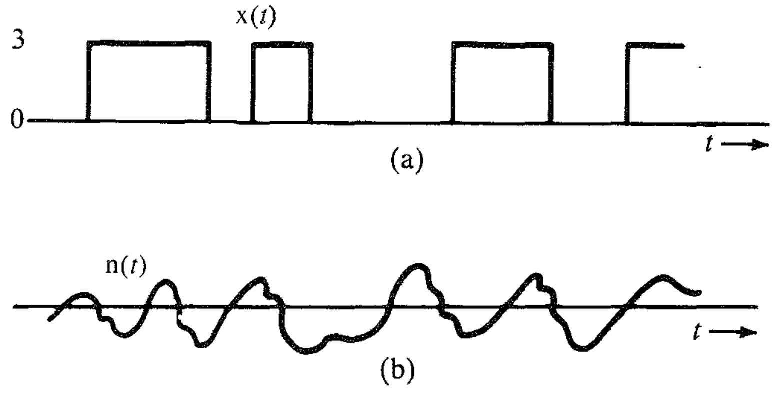


Figure P10.4-1

10.4-2 Repeat Prob. 10.4-1 if the amplitudes 3 and 0 of $\mathrm{x}(t)$ are not equiprobable but $P_{\mathrm{x}}(3) = 0.6$ and $P_{\mathrm{x}}(0) = 0.4$.

10.3-4 If $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are both independent binary signals each taking on values -1 and 1 only with

$$P_{\mathrm{x}}(1) = Q = 1 - P_{\mathrm{x}}(-1)$$

$$P_{\mathrm{y}}(1) = P = 1 - P_{\mathrm{y}}(-1)$$

determine $P_{\mathrm{z}}(z_i)$ where $\mathrm{z} = \mathrm{x} + \mathrm{y}$.

10.4-4 If $\mathrm{z} = \mathrm{x} + \mathrm{y}$, where x and y are independent gaussian RVs with

$$p_{\mathrm{x}}(x) = \frac{1}{\sigma_{\mathrm{x}}\sqrt{2\pi}} e^{-(x-\bar{\mathrm{x}})^2/2\sigma_{\mathrm{x}}^2} \qquad \text{and} \qquad p_{\mathrm{y}}(y) = \frac{1}{\sigma_{\mathrm{y}}\sqrt{2\pi}} e^{-(y-\bar{\mathrm{y}})^2/2\sigma_{\mathrm{y}}^2}$$

then show that z is also Gaussian with

$$\bar{\mathrm{z}} = \bar{\mathrm{x}} + \bar{\mathrm{y}} \qquad \text{and} \qquad \sigma_{\mathrm{z}}^2 = \sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2$$

Hint: Use Eq. (3.43) to convolve $p_{\mathrm{x}}(x)$ and $p_{\mathrm{y}}(y)$. See pair 22 in Table 3.1.

10.5-1 Show that $|\rho_{\mathrm{xy}}| \leq 1$, where ρ_{xy} is the correlation coefficient [Eq. (10.76)] of RVs x and y. *Hint:* For any real number a,

$$\overline{[a(\mathrm{x} - \bar{\mathrm{x}}) - (\mathrm{y} - \bar{\mathrm{y}})]^2} \geq 0$$

The discriminant of this quadratic in a is nonpositive.

10.5-2 Show that if two RVs x and y are related by

$$\mathrm{y} = k_1\mathrm{x} + k_2$$

where k_1 and k_2 are arbitrary constants, the correlation coefficient $\rho_{\mathrm{xy}} = 1$ if k_1 is positive, and $\rho_{\mathrm{xy}} = -1$ if k_1 is negative.

10.5-3 Given $\mathrm{x} = \cos\Theta$ and $\mathrm{y} = \sin\Theta$, where Θ is an RV uniformly distributed in the range $(0, 2\pi)$, show that x and y are uncorrelated but are not independent.

10.6-1 In Example 10.21, design the optimum third-order predictor processor for speech signals and determine the SNR improvement. Values of various correlation coefficients for speech signals are given in Example 10.21.

11 RANDOM PROCESSES

The notion of a random process is an extension of the random variable (RV). Consider, for example, the temperature x of a certain city at noon. The temperature x is an RV and takes on different values every day. To get the complete statistics of x, we need to record values of x at noon over many days (a large number of trials). From this data, we can determine $p_{\mathrm{x}}(x)$, the PDF of the RV x (the temperature at noon).

But the temperature is also a function of time. At 1 P.M., for example, the temperature may have an entirely different distribution from that of the temperature at noon. Thus, the RV x is a function of time and can be expressed as $\mathrm{x}(t)$. An RV that is a function of time* is called a **random process**, or **stochastic process**, which is the subject of this chapter. Thus, a random process is a collection of an infinite number of RVs.

11.1 FROM RANDOM VARIABLE TO RANDOM PROCESS

To specify an RV x, we repeat the experiment a large number of times and from the outcomes determine $p_{\mathrm{x}}(x)$. Similarly, to specify the random process $\mathrm{x}(t)$, we do the same thing for each value of t. To continue with our example of the random process $\mathrm{x}(t)$, the temperature of the city, we need to record daily temperatures for each value of t (for each time of the day). This can be done by recording temperatures at every instant of the day, which gives a waveform $x(t, \zeta_i)$, where ζ_i indicates the day for which the record was taken. We need to repeat this procedure every day for a large number of days. The collection of all possible waveforms is known as the **ensemble** (corresponding to the sample space) of the random process $\mathrm{x}(t)$. A waveform in this collection is a **sample function** (rather than a sample point) of the random process (Fig. 11.1). Sample-function amplitudes at some instant $t = t_1$ are the values taken by the RV $\mathrm{x}(t_1)$ in various trials.

We can view a random process in another way. In the case of an RV, the outcome of each trial of the random experiment is a number. We can view a random process also as the

---

* Actually, to qualify as a random process, x could be a function of any other variable, such as distance. A random process may also be a function of more than one variable.

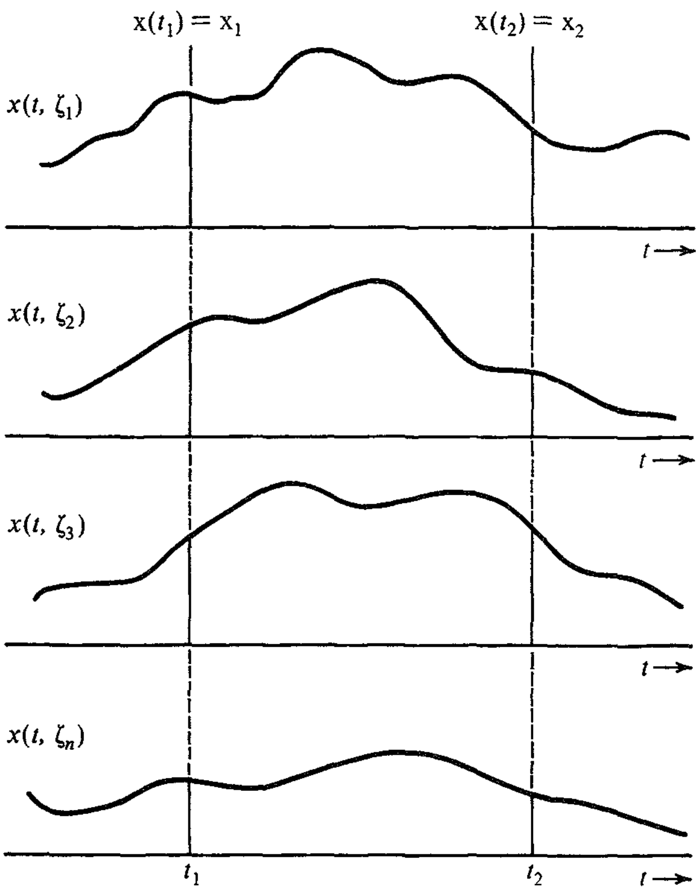


Figure 11.1 Random process to represent the temperature of a city.

outcome of a random experiment, where the outcome of each trial is a waveform (a sample function) that is a function of t. The number of waveforms in an ensemble may be finite or infinite. In the case of the random process $\mathrm{x}(t)$ (the temperature of a city), the ensemble has infinite waveforms. On the other hand, if we consider the output of a binary signal generator (over the period 0 to $10T$), there are at most 2^{10} waveforms in this ensemble (Fig. 11.2).

One fine point that needs clarification is that the waveforms (sample functions) in the ensemble are not random. They are deterministic. Randomness in this situation is associated not with the waveform but with the uncertainty as to which waveform will occur in a given trial. This is completely analogous to the situation of an RV. For example, in the experiment of tossing a coin four times in succession (Example 10.4), 16 possible outcomes exist, all of which are known. The randomness in this situation is associated not with the outcomes but with the uncertainty as to which of the 16 outcomes will occur in a given trial.

Specification of a Random Process

The next important question is how to specify a random process. In some cases, we may be able to describe it analytically. Consider, for instance, a random process described by $\mathrm{x}(t) = A\cos(\omega_c t + \Theta)$, where Θ is an RV uniformly distributed over the range $(0, 2\pi)$. This analytical expression completely describes a random process (and its ensemble). Each sample function is a sinusoid of amplitude A and frequency ω_c. But the phase is random (see Fig. 11.5). It is equally likely to take any value in the range $(0, 2\pi)$.

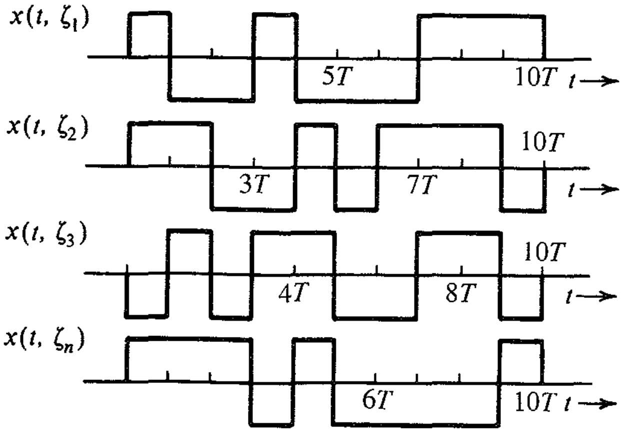


Figure 11.2 Ensemble with a finite number of sample functions.

Unfortunately, it is not always possible to be able to describe a random process analytically. We may have just an ensemble obtained experimentally. The ensemble has the complete information about the random process. From this ensemble, we must find some quantitative measure that will specify or characterize the random process. In this case, we consider the random process as an RV x that is a function of time. Thus, a random process is just a collection of an infinite number of RVs, which are generally dependent. We know that the complete information of several dependent RVs is provided by the joint PDF of those variables. Let x_i represent the RV $\mathrm{x}(t_i)$ generated by the amplitudes of the random process at instant $t = t_i$. Thus, x_1 is the RV generated by the amplitudes at $t = t_1$, and x_2 is the RV generated by the amplitudes at $t = t_2$, and so on, as shown in Fig. 11.1. The n RVs $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_n$ generated by the amplitudes at $t = t_1, t_2, t_3, \ldots, t_n$, respectively, are dependent in general. From this discussion, it follows that the random process is completely described by the joint PDF $p_{\mathrm{x}_1\mathrm{x}_2\ldots\mathrm{x}_n}(x_1, x_2, \ldots, x_n)$ for all n (up to ∞) and for any choice of $t_1, t_2, t_3, \ldots, t_n$. This nth-order joint PDF is also expressed as $p_{\mathrm{x}_1\mathrm{x}_2\ldots\mathrm{x}_n}(x_1, x_2, \ldots, x_n;\ t_1, t_2, \ldots, t_n)$. Determining this PDF (of infinite order) is a formidable task. Fortunately, we shall soon see that when dealing with random processes in conjunction with linear systems, we can get by with only the first- and second-order statistics.

We can always derive a lower order PDF from a higher order PDF by simple integration. For instance,

$$p_{\mathrm{x}_1}(x_1) = \int_{-\infty}^{\infty} p_{\mathrm{x}_1\mathrm{x}_2}(x_1, x_2)\, dx_2$$

Hence, when the nth-order PDF is available, there is no need to specify PDFs of order lower than n.

The mean $\overline{\mathrm{x}(t)}$ of a random process x(t) can be determined from the first-order PDF as

$$\overline{\mathrm{x}(t)} = \int_{-\infty}^{\infty} x p_{\mathrm{x}}(x;\ t)\, dx \tag{11.1}$$

Why Do We Need Ensemble Statistics?

The preceding discussion shows that to specify a random process, we need ensemble statistics. For instance, to determine the PDF $p_{\mathrm{x}_1}(x_1)$, we need to find the amplitudes of all the sample functions at $t = t_1$. This is *ensemble statistics*. In the same way, all possible statistics in the specification of a random process necessitates some kind of ensemble statistics. In deterministic signals, we are used to studying the data of a waveform (or waveforms) as a function of time. Hence, the idea of investigating ensemble statistics makes us feel a bit uncomfortable at first. Theoretically, we may accept it, but does it have any practical significance? How is this concept useful in practice? We shall now answer this question.

To understand the necessity of ensemble statistics, consider the problem of threshold detection in Example 10.13. A **1** is transmitted by $p(t)$ and a **0** is transmitted by $-p(t)$ (polar signaling). The peak pulse amplitude is A_p. When **1** is transmitted, the received sample value is $A_p + n$, where n is the noise. We shall make a decision error if the noise amplitude at the sampling instant t_s is less than $-A_p$. To find this error probability, we repeat the experiment N times ($N \to \infty$), and see how many times the noise amplitude at $t = t_s$ is less than $-A_p$ (Fig. 11.3). This information is precisely one of ensemble statistics of the noise process $n(t)$ at instant t_s.

The importance of ensemble statistics is clear from this example. When we are dealing with a random process or processes, we do not know which sample function will occur in a given trial. Hence, for any specification, characterization, or optimization, we need to average over the entire ensemble. This is the basic physical reason for the appearance of ensemble statistics in random processes.

Autocorrelation Function of a Random Process

One of the most important characteristics of a random process is its **autocorrelation function**, which leads to the spectral information of the random process. The frequency content of a process depends on the rapidity of the amplitude change with time. This can be measured by correlating amplitudes at t_1 and $t_1 + \tau$. The random process $x(t)$ in Fig. 11.4a is a slowly varying process compared to the process $y(t)$ in Fig. 11.4b. For $x(t)$, the amplitudes at t_1 and $t_1 + \tau$ are similar (Fig. 11.4a), that is, have stronger correlation. On the other hand, for $y(t)$, the amplitudes at t_1 and $t_1 + \tau$ have little resemblance (Fig. 11.4b), that is, have weaker correlation. Recall that correlation is a measure of the similarity of two RVs. Hence, we can use correlation

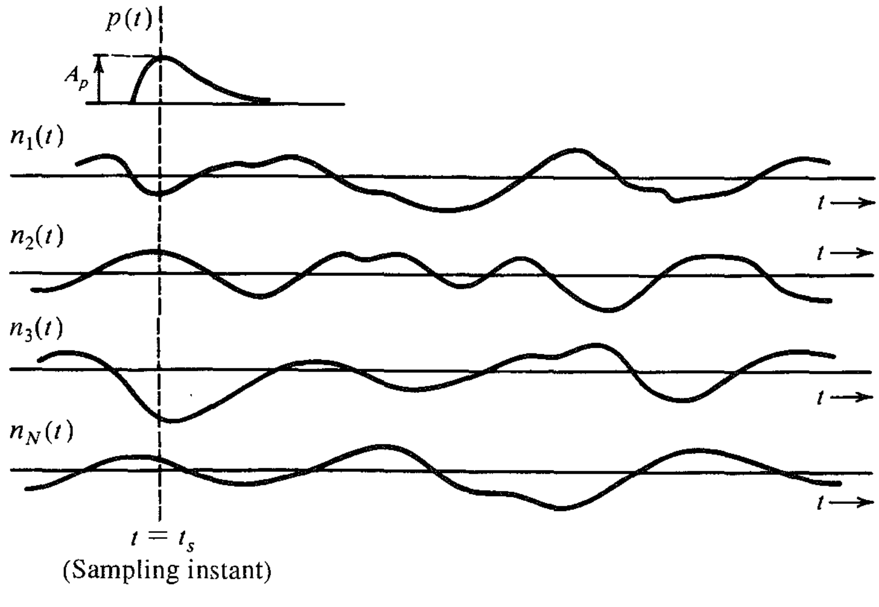


Figure 11.3 A random process to represent a channel noise.

to measure the similarity of amplitudes at t_1 and $t_2 = t_1 + \tau$. If the RVs $\mathrm{x}(t_1)$ and $\mathrm{x}(t_2)$ are denoted by x_1 and x_2, respectively, then for a real random process,* the autocorrelation function $R_{\mathrm{x}}(t_1, t_2)$ is defined as

$$R_{\mathrm{x}}(t_1, t_2) = \overline{\mathrm{x}(t_1)\mathrm{x}(t_2)} = \overline{\mathrm{x}_1\mathrm{x}_2} \tag{11.2a}$$

This is the correlation of RVs $\mathrm{x}(t_1)$ and $\mathrm{x}(t_2)$. It is computed by multiplying amplitudes at t_1 and t_2 of a sample function and then averaging this product over the ensemble. It can be seen that for a small τ, the product x_1x_2 will be positive for most sample functions of $\mathrm{x}(t)$, but the product y_1y_2 will be equally likely to be positive or negative. Hence, $\overline{\mathrm{x}_1\mathrm{x}_2}$ will be larger than $\overline{\mathrm{y}_1\mathrm{y}_2}$. Moreover, x_1 and x_2 will show correlation for considerably larger values of τ, whereas y_1 and y_2 will lose correlation quickly, even for small τ, as shown in Fig. 11.4c.

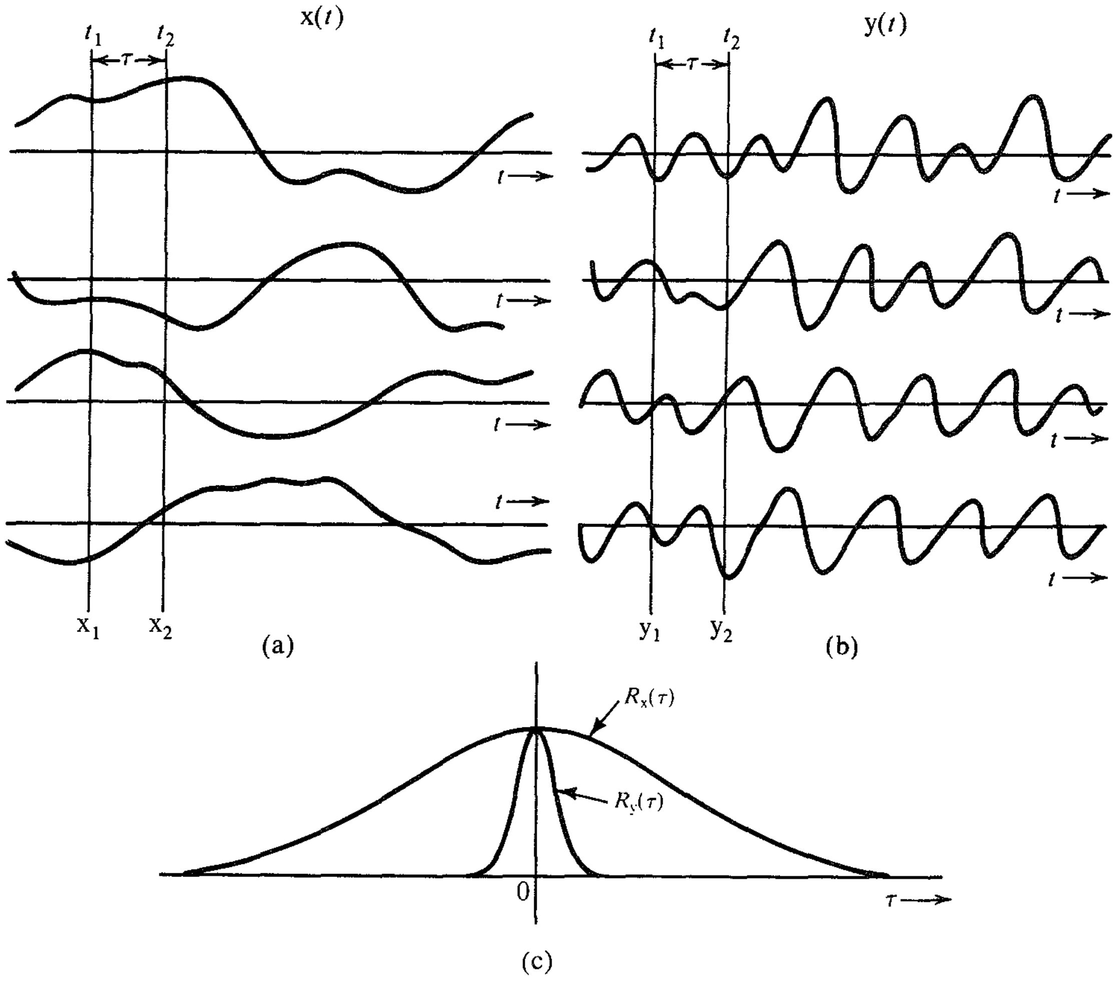


Figure 11.4 Autocorrelation functions for a slowly varying and a rapidly varying random process.

---

* For a complex random process $\mathrm{x}(t)$, the autocorrelation function is defined as

$$R_{\mathrm{x}}(t_1, t_2) = \overline{\mathrm{x}^*(t_1)\mathrm{x}(t_2)}$$

Thus, $R_{\mathrm{x}}(t_1, t_2)$, the autocorrelation function of $\mathrm{x}(t)$, provides valuable information about the frequency content of the process. In fact, we shall show that the PSD of $\mathrm{x}(t)$ is the Fourier transform of its autocorrelation function, given by (for real processes)

$$\begin{aligned} R_{\mathrm{x}}(t_1, t_2) &= \overline{\mathrm{x}_1\mathrm{x}_2} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 p_{\mathrm{x}_1\mathrm{x}_2}(x_1, x_2)\, dx_1 dx_2 \end{aligned} \tag{11.2b}$$

Hence, $R_{\mathrm{x}}(t_1, t_2)$ can be derived from the joint PDF of x_1 and x_2, which is the second-order PDF.

Classification of Random Processes

Random processes may be classified in the following broad categories.

Stationary and Nonstationary Random Processes: A random process whose statistical characteristics do not change with time is classified as a **stationary random process.** For a stationary process, we can say that a shift of time origin will be impossible to detect; the process will appear to be the same. Suppose we determine $p_{\mathrm{x}}(x;\ t_1)$, then shift the origin by t_0, and again determine $p_{\mathrm{x}}(x;\ t_1)$. The instant t_1 in the new frame of reference is $t_2 = t_1 + t_0$ in the old frame of reference. Hence, the PDFs of x at t_1 and $t_2 = t_1 + t_0$ must be the same, that is, $p_{\mathrm{x}}(x;\ t_1)$ and $p_{\mathrm{x}}(x;\ t_2)$ must be identical for a stationary random process. This is possible only if $p_{\mathrm{x}}(x;\ t)$ is independent of t. Thus, the first-order density of a stationary random process can be expressed as

$$p_{\mathrm{x}}(x;\ t) = p_{\mathrm{x}}(x)$$

Similarly, for a stationary random process the autocorrelation function $R_{\mathrm{x}}(t_1, t_2)$ must depend on t_1 and t_2 only through the difference $t_2 - t_1$. If not, we could determine a unique time origin. Hence, for a real stationary process,

$$R_{\mathrm{x}}(t_1, t_2) = R_{\mathrm{x}}(t_2 - t_1)$$

Therefore,

$$R_{\mathrm{x}}(\tau) = \overline{\mathrm{x}(t)\mathrm{x}(t+\tau)} \tag{11.3}$$

For a stationary process, the joint PDF for x_1 and x_2 must also depend only on $t_2 - t_1$. Similarly, higher order PDFs, such as $p_{\mathrm{x}_1\mathrm{x}_2\ldots\mathrm{x}_n}(x_1, x_2, \ldots, x_n)$ where $\mathrm{x}_i = \mathrm{x}(t_i)$, are all independent of the choice of origin.

The random process $\mathrm{x}(t)$ representing the temperature of a city is an example of a nonstationary random process, because the temperature statistics (mean value, for example) depend on the time of the day. On the other hand, the noise process in Fig. 11.3 is stationary, because its statistics (the mean and the mean square values, for example) do not change with time. In general, it is not easy to determine whether a process is stationary, because it involves investigation of the nth-order ($n = \infty$) statistics. In practice, we can ascertain stationarity if there is no change in the signal-generating mechanism. Such is the case for the noise process in Fig. 11.3.

Wide-Sense (or Weakly) Stationary Processes: A process may not be stationary in the *strict* sense, as discussed in the last subsection, yet it may have a mean value and an autocorrelation function that are independent of the shift of time origin. This means

$$\overline{\mathrm{x}(t)} = \text{constant}$$

and

$$R_{\mathrm{x}}(t_1,\ t_2) = R_{\mathrm{x}}(\tau) \qquad \tau = t_2 - t_1 \tag{11.4}$$

Such a process is known as a **wide-sense stationary**, or **weakly stationary, process.** Note that stationarity is a much stronger condition than wide-sense stationarity. All stationary processes are wide-sense stationary, but the converse is not necessarily true.

Just as no sinusoidal signal exists in actual practice, no truly stationary process can occur in real life. All processes in practice are nonstationary because they must begin at some finite time and must terminate at some finite time. A truly stationary process must start at $t = -\infty$ and go on forever. Many processes can be considered stationary for the time interval of interest, however, and the stationarity assumption allows a manageable mathematical model. The use of a stationary model is analogous to the use of a sinusoidal model in deterministic analysis.

EXAMPLE 11.1 Show that the random process

$$\mathrm{x}(t) = A \cos(\omega_c t + \Theta)$$

where Θ is an RV uniformly distributed in the range $(0,\ 2\pi)$, is a wide-sense stationary process.

The ensemble (Fig. 11.5) consists of sinusoids of constant amplitude A and constant frequency ω_c, but the phase Θ is random. For any sample function, the phase is equally likely to have any value in the range $(0,\ 2\pi)$. Because Θ is an RV uniformly distributed over the range $(0,\ 2\pi)$, one can determine[1] $p_{\mathrm{x}}(x,\ t)$ and, hence, $\overline{\mathrm{x}(t)}$, as in Eq. (11.1). For this particular case, however, $\overline{\mathrm{x}(t)}$ can be determined directly as follows:

$$\overline{\mathrm{x}(t)} = \overline{A \cos(\omega_c t + \Theta)} = A\, \overline{\cos(\omega_c t + \Theta)}$$

Because $\cos(\omega_c t + \Theta)$ is a function of an RV Θ, we have [see Eq. (10.57b)]

$$\overline{\cos(\omega_c t + \Theta)} = \int_0^{2\pi} \cos(\omega_c t + \theta) p_\Theta(\theta)\, d\theta$$

Because $p_\Theta(\theta) = 1/2\pi$ over $(0,\ 2\pi)$ and 0 outside this range,

$$\overline{\cos(\omega_c t + \Theta)} = \frac{1}{2\pi} \int_0^{2\pi} \cos(\omega_c t + \theta)\, d\theta = 0$$

Hence,

$$\overline{\mathrm{x}(t)} = 0 \tag{11.5a}$$

Thus, the ensemble mean of sample-function amplitudes at any instant t is zero.

The autocorrelation function $R_{\mathrm{x}}(t_1,\ t_2)$ for this process also can be determined directly from Eq. (11.2a),

$$\begin{aligned} R_x(t_1,\ t_2) &= \overline{A^2 \cos(\omega_c t_1 + \Theta)\cos(\omega_c t_2 + \Theta)} \\ &= A^2\, \overline{\cos(\omega_c t_1 + \Theta)\cos(\omega_c t_2 + \Theta)} \\ &= \frac{A^2}{2}\left\{\overline{\cos[\omega_c(t_2 - t_1)]} + \overline{\cos[\omega_c(t_2 + t_1) + 2\Theta]}\right\} \end{aligned}$$

The first term on the right-hand side contains no RV. Hence, $\overline{\cos[\omega_c(t_2 - t_1)]}$ is $\cos[\omega_c(t_2 - t_1)]$ itself. The second term is a function of the RV Θ, and its mean is

$$\overline{\cos[\omega_c(t_2 + t_1) + 2\Theta]} = \frac{1}{2\pi}\int_0^{2\pi} \cos[\omega_c(t_2 + t_1) + 2\theta]\, d\theta = 0$$

Hence,

$$R_x(t_1,\ t_2) = \frac{A^2}{2}\cos[\omega_c(t_2 - t_1)] \tag{11.5b}$$

or

$$R_x(\tau) = \frac{A^2}{2}\cos\omega_c\tau \qquad \tau = t_2 - t_1 \tag{11.5c}$$

From Eqs. (11.5a,b) it is clear that $x(t)$ is a wide-sense stationary process.

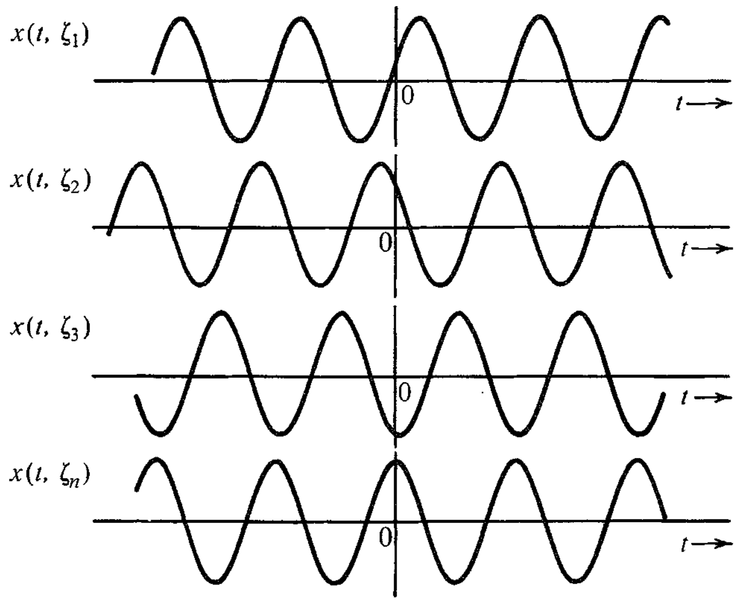


Figure 11.5 Ensemble for the random process $A\cos(\omega_c t + \Theta)$.

Ergodic Processes: We have studied the mean and the autocorrelation function of a random process. These are ensemble averages. For example, $\overline{x(t)}$ is the ensemble average of sample-function amplitudes at t, and $R_x(t_1,\ t_2) = \overline{x_1 x_2}$ is the ensemble average of the product of sample-function amplitudes $x(t_1)$ and $x(t_2)$.

We can also define time averages for each sample function. For example, a time mean $\widetilde{x(t)}$ of a sample function $x(t)$ is*

$$\widetilde{x(t)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)\,dt \tag{11.6a}$$

Similarly, the time-autocorrelation function $\mathcal{R}_{\mathrm{x}}(\tau)$ defined in Eq. (3.82a) is

$$\mathcal{R}_{\mathrm{x}}(\tau) = \widetilde{x(t)x(t+\tau)} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)\,dt \tag{11.6b}$$

For **ergodic processes**, ensemble averages are equal to the time averages of any sample function. Thus, for an ergodic process $\mathrm{x}(t)$,

$$\overline{\mathrm{x}(t)} = \widetilde{x(t)} \tag{11.7a}$$

$$R_{\mathrm{x}}(\tau) = \mathcal{R}_{\mathrm{x}}(\tau) \tag{11.7b}$$

These are just two of the many possible averages. For an ergodic process, all possible ensemble averages are equal to the corresponding time averages of one of its sample functions. Because a time average cannot be a function of time, it is evident that an ergodic process is necessarily a stationary process; but the converse is not true (see Fig. 11.6).

It is difficult to test whether a process is ergodic or not, because we must test all possible orders of time and ensemble averages. Nevertheless, in practice many of the stationary processes are ergodic with respect to at least second-order averages, such as the mean and the autocorrelation. For the process in Example 11.1 (Fig. 11.5), we can show that $\widetilde{x(t)} = 0$ and $\mathcal{R}_{\mathrm{x}}(\tau) = (A^2/2)\cos\omega_c\tau$ (see Problem 3.8-1). Therefore, this process is ergodic at least with respect to the first- and second-order averages.

The ergodicity concept can be explained by a simple example of traffic lights in a city. Suppose the city is well planned, with all its streets in E–W and N–S directions only and with traffic lights at each intersection. Assume that each light stays green for 0.75 second in the E–W direction and 0.25 second in the N–S direction and that switching of any light is independent of the other lights. For the sake of simplicity, we ignore the orange light.

If we consider a certain person driving a car arriving at any traffic light randomly in the E–W direction, the probability that the person will have a green light is 0.75; that is, on the average, 75% of the time the person will observe a green light. On the other hand, if we consider a large number of drivers arriving at a traffic light in the E–W direction at some instant

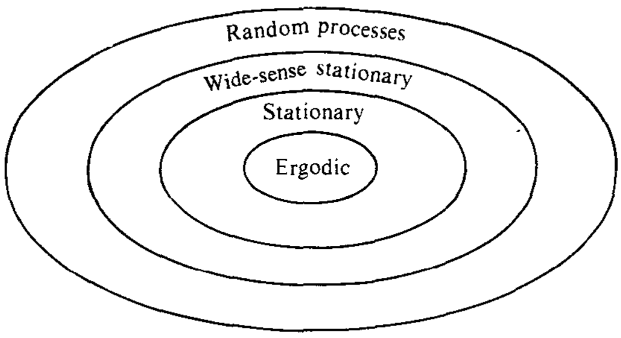


Figure 11.6 Classification of random processes.

---

* Here a sample function $x(t, \zeta_i)$ is represented by $x(t)$ for convenience.

t, then 75% of the drivers will have a green light, and the remaining 25% will have a red light. Thus, the experience of a single driver arriving randomly many times at a traffic light will contain the same statistical information (sample-function statistics) as that of a large number of drivers arriving simultaneously at various traffic lights (ensemble statistics) at one instant.

The ergodicity notion is extremely important, because we do not have a large number of sample functions available in practice from which to compute ensemble averages. If the process is known to be ergodic, then we need only one sample function to compute ensemble averages. As mentioned earlier, many of the stationary processes encountered in practice are ergodic with respect to at least second-order averages. As we shall see in dealing with stationary processes in conjunction with linear systems, we need only the first- and second-order averages. This means that in most cases we can get by with a single sample function.

11.2 POWER SPECTRAL DENSITY OF A RANDOM PROCESS

An electrical engineer instinctively thinks of signals and linear systems in terms of their frequency-domain descriptions. Linear systems are characterized by their frequency response (the transfer function), and signals are expressed in terms of the relative amplitudes and phases of their frequency components (the Fourier transform). From a knowledge of the input spectrum and transfer function, the response of a linear system to a given signal can be obtained in terms of the frequency content of that signal. This is an important procedure for deterministic signals. We may wonder if similar methods may be found for random processes. Ideally, all the sample functions of a random process are assumed to exist over the entire time interval $(-\infty, \infty)$ and thus, are power signals.* We therefore inquire about the existence of a power spectral density (PSD). Superficially, the concept of a PSD of a random process may appear ridiculous for the following reasons. In the first place, we may not be able to describe a sample function analytically. Second, for a given process, every sample function may be different from another one. Hence, even if a PSD does exist for each sample function, it may be different for different sample functions. Fortunately, both problems can be neatly resolved, and it is possible to define a meaningful PSD for a stationary (at least in the wide sense) random process. For nonstationary processes, the PSD does not exist.

Whenever randomness is involved, our inquiries can at best provide answers in terms of averages. When tossing a coin, for instance, the most we can say about the outcome is that on the average we will obtain heads in about half the trials and tails in the remaining half of the trials. For random signals or RVs, we do not have enough information to predict the outcome with certainty, and we must accept answers in terms of averages. It is not possible to transcend this limit of knowledge because of the fundamental ignorance of the process. It seems reasonable to define the PSD of a random process as a weighted mean of the PSDs of all sample functions. This is the only sensible solution, because we do not know exactly which of the sample functions may occur in a given trial. We must be prepared for any sample function. Consider, for example, the problem of filtering a certain random process. We would not want to design a filter with respect to any one particular sample function because any of the sample functions in the ensemble may be present at the input. A sensible approach is to design the filter with respect to the mean parameters of the input process. In designing a system to perform certain operations, one must design it with respect to the whole ensemble. We are therefore justified in defining the PSD $S_x(\omega)$ of a random process $x(t)$ as the ensemble average of the PSDs of all sample functions. Thus [see Eq. (3.79)],

---

* As we shall soon see, for the PSD to exist, the process must be stationary (at least in the wide sense). Stationary processes, because their statistics do not change with time, are power signals.

$$S_x(\omega) = \lim_{T\to\infty} \overline{\left[\frac{|X_T(\omega)|^2}{T}\right]} \tag{11.8a}$$

where $X_T(\omega)$ is the Fourier transform of the truncated random process $x(t)\,\text{rect}\,(t/T)$ and the bar represents ensemble average. Note that ensemble averaging is done before the limiting operation. We shall now show that the PSD as defined in Eq. (11.8a) is the Fourier transform of the autocorrelation function $R_x(\tau)$ of the process $x(t)$; that is,

$$R_x(\tau) \Longleftrightarrow S_x(\omega) \tag{11.8b}$$

This can be proved as follows:

$$X_T(\omega) = \int_{-\infty}^{\infty} x_T(t)e^{-j\omega t}\,dt = \int_{-T/2}^{T/2} x(t)e^{-j\omega t}\,dt \tag{11.9}$$

Thus, for real $x(t)$,

$$\begin{aligned}
|X_T(\omega)|^2 &= X_T(-\omega)X_T(\omega) \\
&= \int_{-T/2}^{T/2} x(t_1)e^{j\omega t_1}\,dt_1 \int_{-T/2}^{T/2} x(t_2)e^{-j\omega t_2}\,dt_2 \\
&= \int_{-T/2}^{T/2}\int_{-T/2}^{T/2} x(t_1)x(t_2)e^{-j\omega(t_2-t_1)}\,dt_1dt_2
\end{aligned}$$

and

$$\begin{aligned}
S_x(\omega) &= \lim_{T\to\infty} \overline{\left[\frac{|X_T(\omega)|^2}{T}\right]} \\
&= \lim_{T\to\infty} \overline{\left[\frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} x(t_1)x(t_2)e^{-j\omega(t_2-t_1)}\,dt_1dt_2\right]}
\end{aligned} \tag{11.10}$$

Interchanging the operation of integration and ensemble averaging,* we get

$$\begin{aligned}
S_x(\omega) &= \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} \overline{x(t_1)x(t_2)}e^{-j\omega(t_2-t_1)}\,dt_1dt_2 \\
&= \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} R_x(t_2-t_1)e^{-j\omega(t_2-t_1)}\,dt_1dt_2
\end{aligned}$$

Here we are assuming that the process $x(t)$ is at least wide-sense stationary, so that $\overline{x(t_1)x(t_2)} = R_x(t_2 - t_1)$. For convenience, let

---

* The operation of ensemble averaging is also an operation of integration. Hence, interchanging integration with ensemble averaging is equivalent to interchanging the order of integration.

$$R_x(t_2 - t_1)e^{-j\omega(t_2-t_1)} = \varphi(t_2 - t_1) \tag{11.11}$$

Then,

$$S_x(\omega) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \varphi(t_2 - t_1)\, dt_1 dt_2 \tag{11.12}$$

The integral on the right-hand side is a double integral over the range $(-T/2,\ T/2)$ for each of the variables t_1 and t_2. The square region of integration in the t_1–t_2 plane is shown in Fig. 11.7. The integral in Eq. (11.12) is a volume under the surface $\varphi(t_2 - t_1)$ over the square region in Fig. 11.7. The double integral in Eq. (11.12) can be converted to a single integral by observing that $\varphi(t_2 - t_1)$ is constant along any line $t_2 - t_1 = \tau$ (a constant) in the t_1–t_2 plane (Fig. 11.7).

Let us consider two such lines, $t_2 - t_1 = \tau$ and $t_2 - t_1 = \tau + \Delta\tau$. If $\Delta\tau \to 0$, $\varphi(t_2 - t_1) \simeq \varphi(\tau)$ over the shaded region whose area is $(T - \tau)\,\Delta\tau$. Hence, the volume under the surface $\varphi(t_2 - t_1)$ over the shaded region is $\varphi(\tau)(T - \tau)\,\Delta\tau$. If τ were negative, the volume would be $\varphi(\tau)(T + \tau)\,\Delta\tau$. Hence, in general, the volume over the shaded region is $\varphi(\tau)(T - |\tau|)\,\Delta\tau$. The desired volume over the square region in Fig. 11.7 is the sum of the volumes over the shaded strips and is obtained by integrating $\varphi(\tau)(T - |\tau|)$ over the range of τ, which is $(-T,\ T)$ (see Fig. 11.7). Hence,

$$\begin{aligned} S_x(\omega) &= \lim_{T\to\infty} \frac{1}{T} \int_{-T}^{T} \varphi(\tau)(T - |\tau|)\, d\tau \\ &= \lim_{T\to\infty} \int_{-T/2}^{T/2} \varphi(\tau)\left(1 - \frac{|\tau|}{T}\right) d\tau \\ &= \int_{-\infty}^{\infty} \varphi(\tau)\, d\tau \end{aligned}$$

provided $\int_{-\infty}^{\infty} |\tau|\varphi(\tau)\, d\tau$ is bounded. Substituting Eq. (11.11) into this equation, we have

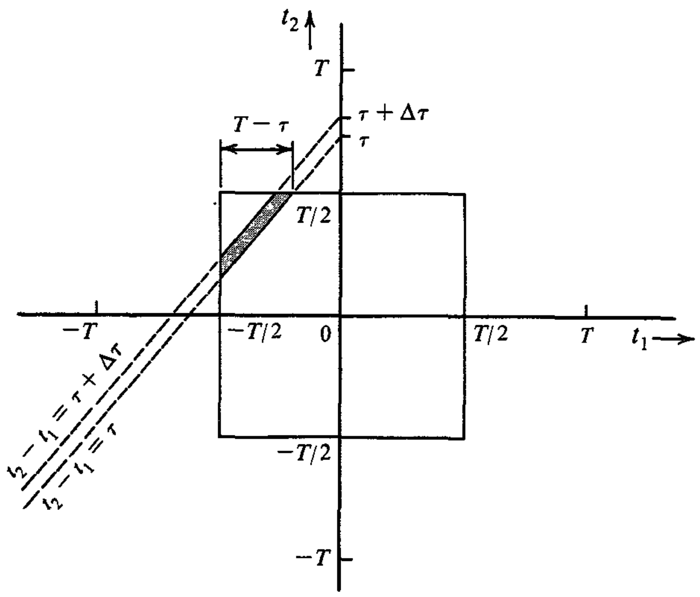


Figure 11.7 Derivation of Wiener-Khinchine theorem.

$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau) e^{-j\omega\tau}\, d\tau \tag{11.13}$$

provided $\int_{-\infty}^{\infty} |\tau| R_x(\tau) e^{-j\omega\tau}\, d\tau$ is bounded. Thus, the PSD of a wide-sense stationary random process is the Fourier transform of its autocorrelation function,*

$$R_x(\tau) \Longleftrightarrow S_x(\omega) \tag{11.14}$$

This is the well-known Wiener-Khinchine theorem.

From the discussion thus far, the autocorrelaton function emerges as one of the most significant entities in the spectral analysis of a random process. Earlier we showed heuristically how the autocorrelation function is connected with the frequency content of a random process.

The autocorrelation function $R_x(\tau)$ for real processes is an even function of τ. This can be proved in two ways. First, because $|X_T(\omega)|^2 = \left|X_T(\omega) X_T^*(\omega)\right| = |X_T(\omega) X_T(-\omega)|$ is an even function of ω, $S_x(\omega)$ is also an even function of ω, and $R_x(\tau)$, its inverse transform, is also an even function of τ (see Prob. 3.1-1). Alternately, we may argue that

$$R_x(\tau) = \overline{x(t)x(t+\tau)} \qquad \text{and} \qquad R_x(-\tau) = \overline{x(t)x(t-\tau)}$$

Letting $t - \tau = \sigma$, we have

$$R_x(-\tau) = \overline{x(\sigma)x(\sigma+\tau)} = R_x(\tau) \tag{11.15}$$

The PSD $S_x(\omega)$ is also a real and even function of ω.

The mean square value $\overline{x^2(t)}$ of the random process $x(t)$ is $R_x(0)$,

$$R_x(0) = \overline{x(t)x(t)} = \overline{x^2(t)} = \overline{x^2} \tag{11.16}$$

The mean square value $\overline{x^2}$ is not the time mean square of a sample function but the ensemble average of the amplitude squares of sample functions at any instant t.

The Power of a Random Process

The power P_x (average power) of a wide-sense random process $x(t)$ is its mean square value $\overline{x^2}$. From Eq. (11.14),

$$R_x(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega) e^{j\omega\tau}\, d\omega$$

Hence, from Eq. (11.16),

$$P_x = \overline{x^2} = R_x(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega)\, d\omega \tag{11.17a}$$

$$\frac{d(2\pi f)}{df} = 2\pi$$

Because $S_x(\omega)$ is an even function of ω, we have

$$P_x = \overline{x^2} = 2 \int_0^{\infty} S_x(\omega)\, df \qquad f = \frac{\omega}{2\pi} \tag{11.17b}$$

where f is the frequency in hertz. This is the same relationship as that derived for deterministic signals in Chapter 3 [Eq. (3.81)]. The power P_x is the area under the PSD. Also, $P_x = \overline{x^2}$ is the ensemble mean of the square amplitudes of the sample functions at any instant.

---

* It can be shown that Eq. (11.13) holds also for complex random processes, for which we define $R_x(\tau) = \overline{x^*(t)x(t+\tau)}$.

It is helpful to repeat here, once again, that the PSD does not exist for processes that are not wide-sense stationary. Hence, in our future discussion, random processes will be assumed to be at least wide-sense stationary, unless specifically stated otherwise.

EXAMPLE 11.2 Determine the autocorrelation function $R_x(\tau)$ and the power P_x of a low-pass random process with a white-noise PSD $S_x(\omega) = \mathcal{N}/2$ (Fig. 11.8a).

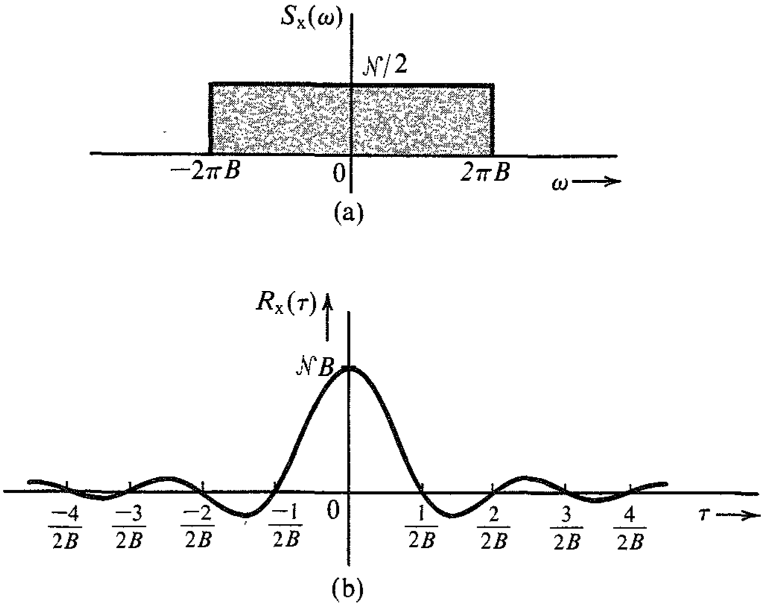


Figure 11.8 Bandpass white-noise PSD and its autocorrelation function.

We have

$$S_x(\omega) = \frac{\mathcal{N}}{2}\operatorname{rect}\left(\frac{\omega}{4\pi B}\right) \tag{11.18a}$$

Hence, from Table 3.1 (pair 17),

$$R_x(\tau) = \mathcal{N}B \operatorname{sinc}(2\pi B\tau) \tag{11.18b}$$

This is shown in Fig. 11.8b. Also,

$$P_x = \overline{x^2} = R_x(0) = \mathcal{N}B \tag{11.18c}$$

Alternately,

$$\begin{aligned} P_x &= 2\int_0^\infty S_x(\omega)\,df \\ &= 2\int_0^B \frac{\mathcal{N}}{2}\,df \\ &= \mathcal{N}B \end{aligned} \tag{11.18d}$$

EXAMPLE 11.3 Determine the PSD and the mean square value of a random process

$$\mathrm{x}(t) = A\cos(\omega_c t + \Theta) \tag{11.19a}$$

where Θ is an RV uniformly distributed over $(0,\ 2\pi)$.

For this case $R_\mathrm{x}(\tau)$ is already determined [Eq. (11.5c)].,

$$R_\mathrm{x}(\tau) = \frac{A^2}{2}\cos\omega_c\tau \tag{11.19b}$$

Hence,

$$S_\mathrm{x}(\omega) = \frac{\pi A^2}{2}[\delta(\omega+\omega_c) + \delta(\omega-\omega_c)] \tag{11.19c}$$

$$P_\mathrm{x} = \overline{\mathrm{x}^2} = R_\mathrm{x}(0) = \frac{A^2}{2} \tag{11.19d}$$

Thus, the power, or the mean square value, of the process $\mathrm{x}(t) = A\cos(\omega_c t + \Theta)$ is $A^2/2$. The power P_x can also be obtained by integrating $S_\mathrm{x}(\omega)$ with respect to ω (or f).

EXAMPLE 11.4 **Amplitude Modulation**

Determine the autocorrelation function and the PSD of the DSB-SC-modulated process $\mathrm{m}(t)\cos(\omega_c t + \Theta)$, where $\mathrm{m}(t)$ is a wide-sense stationary random process, and Θ is an RV uniformly distributed over $(0, 2\pi)$ and independent of $\mathrm{m}(t)$.

Let

$$\varphi(t) = \mathrm{m}(t)\cos(\omega_c t + \Theta)$$

Then

$$R_\varphi(\tau) = \overline{\{\mathrm{m}(t)\cos(\omega_c t + \Theta)][\mathrm{m}(t+\tau)\cos[\omega_c(t+\tau) + \Theta]\}}$$

Because $\mathrm{m}(t)$ and Θ are independent, we can write [see Eqs. (10.60b) and (11.5c)]

$$\begin{aligned} R_\varphi(\tau) &= \overline{\mathrm{m}(t)\mathrm{m}(t+\tau)}\ \overline{\cos(\omega_c t + \Theta)\cos[\omega_c(t+\tau) + \Theta]} \\ &= \frac{1}{2}R_\mathrm{m}(\tau)\cos\omega_c\tau \end{aligned} \tag{11.20a}$$

Consequently,*

$$S_\varphi(\omega) = \frac{1}{4}[S_\mathrm{m}(\omega+\omega_c) + S_\mathrm{m}(\omega-\omega_c)] \tag{11.20b}$$

From Eq. (11.20a) it follows that

$$\overline{\varphi^2(t)} = R_\varphi(0) = \frac{1}{2}R_\mathrm{m}(0) = \frac{1}{2}\overline{\mathrm{m}^2(t)} \tag{11.20c}$$

Hence, the power of the DSB-SC-modulated signal is half the power of the modulating signal. We derived the same result earlier [Eq. (3.92)] for deterministic signals.

---

* We obtain the same result even if $\varphi(t) = \mathrm{m}(t)\sin(\omega_c t + \Theta)$.

EXAMPLE 11.5 Random Binary Process

In this example we shall consider a random binary process for which a typical sample function is shown in Fig. 11.9a. The signal can assume only two states (values), 1 or −1, with equal probability. The transition from one state to another can take place only at node points, which occur every T_b seconds. The probability of a transition from one state to the other is 0.5. The first node is equally likely to be situated at any instant within the interval 0 to T_b from the origin.* The amplitudes at t represent RV x_1, and those at $t + \tau$ represent RV x_2. Note that x_1 and x_2 are discrete and each can assume only two values, −1 and 1. Hence,

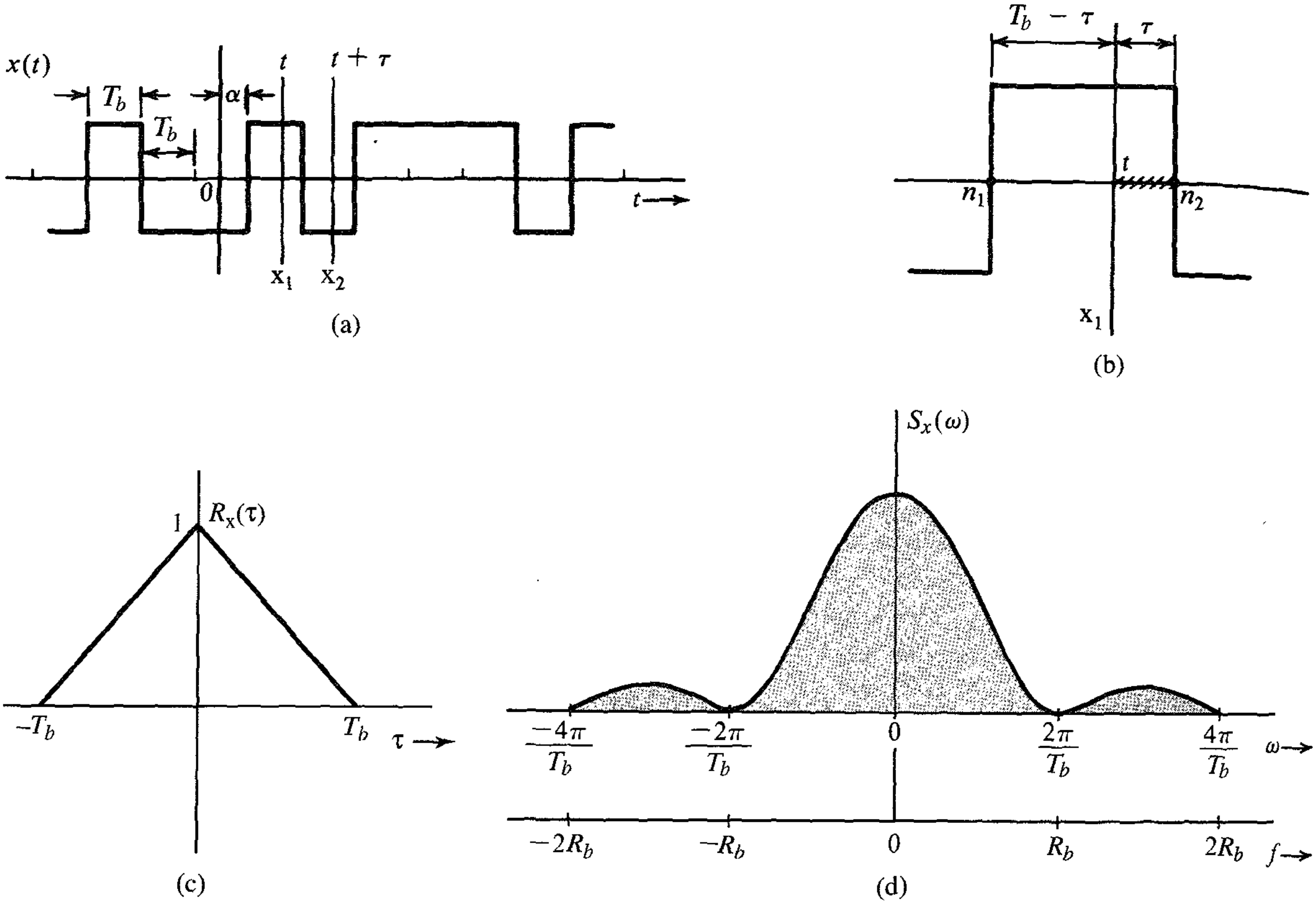


Figure 11.9 Derivation of autocorrelation function and PSD of a random binary process.

$$R_x(\tau) = \overline{x_1 x_2} = \sum_{x_1}\sum_{x_2} x_1 x_2 P_{x_1x_2}(x_1,\, x_2)$$

$$= P_{x_1x_2}(1,\, 1) + P_{x_1x_2}(-1,\, -1) - P_{x_1x_2}(-1,\, 1) - P_{x_1x_2}(1,\, -1) \qquad (11.21a)$$

---

* Analytically, we can represent $x(t)$ as

$$x(t) = \sum_n a_n p(t - nT_b - \alpha)$$

where α is an RV uniformly distributed over the range $(0,\, T_b)$ and $p(t)$ is the basic pulse (in this case $\text{rect}\,[(t - T_b/2)/T_b]$). Note that α is the distance of the first node from the origin, and it varies randomly from sample function to sample function. In addition, a_n is random, taking values 1 or −1 with equal probability.

By symmetry, the first two terms and the last two terms on the right-hand side are equal. Therefore,

$$R_{\mathrm{x}}(\tau) = 2[P_{\mathrm{x}_1\mathrm{x}_2}(1,\ 1) - P_{\mathrm{x}_1\mathrm{x}_2}(1,\ -1)] \tag{11.21b}$$

Using Bayes' rule, we have

$$\begin{aligned} R_{\mathrm{x}}(\tau) &= 2P_{\mathrm{x}_1}(1)[P_{\mathrm{x}_2|\mathrm{x}_1}(1|1) - P_{\mathrm{x}_2|\mathrm{x}_1}(-1|1)] \\ &= P_{\mathrm{x}_2|\mathrm{x}_1}(1|1) - P_{\mathrm{x}_2|\mathrm{x}_1}(-1|1) \end{aligned} \tag{11.21c}$$

Moreover,

$$P_{\mathrm{x}_2|\mathrm{x}_1}(1|1) = 1 - P_{\mathrm{x}_2|\mathrm{x}_1}(-1|1)$$

Hence,

$$R_{\mathrm{x}}(\tau) = 1 - 2P_{\mathrm{x}_2|\mathrm{x}_1}(-1|1)$$

It is helpful to compute $R_{\mathrm{x}}(\tau)$ for small values of τ first. Let us consider the case $\tau < T_b$, where, at most, one node is in the interval t to $t + \tau$. In this case, the event $\mathrm{x}_2 = -1$ given $\mathrm{x}_1 = 1$ is a joint event AB, where the event A is "a node in the interval $(t,\ t + \tau)$" and B is "the state change at this node." Because A and B are independent events,

$$\begin{aligned} P_{\mathrm{x}_2|\mathrm{x}_1}(-1|1) &= P(\text{a node lies in } t \text{ to } t + \tau)P(\text{state change}) \\ &= \tfrac{1}{2}P(\text{a node lies in } t \text{ to } t + \tau) \end{aligned}$$

Figure 11.9b shows adjacent nodes n_1 and n_2, between which t lies. We mark off the interval τ from the node n_2. If t lies anywhere in this interval (hatched area), the node n_2 lies within t and $t + \tau$. But because the instant t is chosen arbitrarily between nodes n_1 and n_2, it is equally likely to be at any instant over the T_b seconds between n_1 and n_2, and the probability that t lies in the shaded interval is simply τ/T_b. Therefore,

$$P_{\mathrm{x}_2|\mathrm{x}_1}(-1|1) = \frac{1}{2}\left(\frac{\tau}{T_b}\right) \tag{11.22}$$

and

$$R_{\mathrm{x}}(\tau) = 1 - \frac{\tau}{T_b} \qquad \tau < T_b \tag{11.23}$$

Because $R_{\mathrm{x}}(\tau)$ is an even function of τ, we have

$$R_{\mathrm{x}}(\tau) = 1 - \frac{|\tau|}{T_b} \qquad |\tau| < T_b \tag{11.24}$$

Next, consider the range $\tau > T_b$. In this case at least one node lies in the interval t to $t + \tau$. Hence, x_1 and x_2 become independent, and

$$R_{\mathrm{x}}(\tau) = \overline{\mathrm{x}_1\mathrm{x}_2} = \overline{\mathrm{x}_1}\,\overline{\mathrm{x}_2} = 0 \qquad \tau > T_b$$

where, by inspection, we observe that $\bar{\mathrm{x}}_1 = \bar{\mathrm{x}}_2 = 0$ (Fig. 11.9a). This result can also be obtained by observing that for $|\tau| > T_b$, x_1 and x_2 are independent, and it is equally likely that $x_2 = 1$ or -1 given that $x_1 = 1$ (or -1). Hence, all four probabilities in Eq. (11.21a) are equal to 1/4, and

$$R_{\mathrm{x}}(\tau) = 0 \qquad \tau > T_b$$

Therefore,

$$R_{\mathrm{x}}(\tau) = \begin{cases} 1 - |\tau|/T_b & |\tau| < T_b \\ 0 & |\tau| > T_b \end{cases} \tag{11.25a}$$

and

$$S_{\mathrm{x}}(\omega) = T_b \operatorname{sinc}^2\left(\frac{\omega T_b}{2}\right) \tag{11.25b}$$

The autocorrelation function and the PSD of this process are shown in Fig. 11.9c and d. Observe that $\overline{\mathrm{x}^2} = R_{\mathrm{x}}(0) = 1$, as expected.

Let us now consider a more general case of the pulse train $\mathrm{y}(t)$, discussed in Sec. 7.2 (Fig. 7.3). From the knowledge of the PSD of this train, we can derive the PSD of on-off, polar, bipolar, duobinary, split-phase, and many more important digital signals.

EXAMPLE 11.6 Random PAM Pulse Train

Digital data is transmitted using a basic pulse $p(t)$, as shown in Fig. 11.10a. The successive pulses are separated by T_b seconds, and the kth pulse is $\mathrm{a}_k p(t)$, where a_k is an RV. The distance α of the first pulse (corresponding to $k = 0$) from the origin is equally likely to be any value in the range $(0,\ T_b)$. Find the autocorrelation function and the PSD of such a random pulse train $\mathrm{y}(t)$ whose sample function is shown in Fig. 11.10b. The random process $\mathrm{y}(t)$ can be described as

$$\mathrm{y}(t) = \sum_{k=-\infty}^{\infty} \mathrm{a}_k p(t - kT_b - \alpha)$$

where α, is an RV uniformly distributed in the interval $(0,\ T_b)$. Thus, α is different for each sample function. Note that $p(\alpha) = 1/T_b$ over the interval $(0,\ T_b)$, and is zero everywhere else.* It can be shown that $\overline{\mathrm{y}(t)} = (\overline{\mathrm{a}_k}/T_b)\int_{-\infty}^{\infty} p(t)\,dt$ is a constant.†

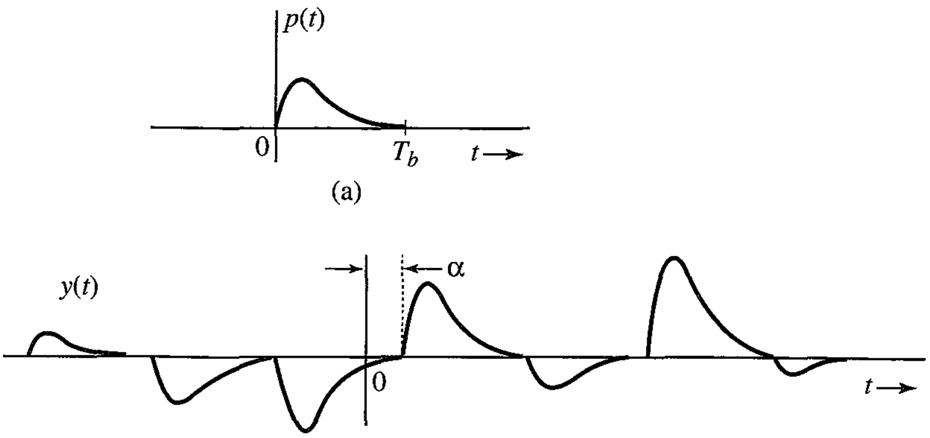
We have the expression

$$\begin{aligned} R_{\mathrm{y}}(\tau) &= \overline{\mathrm{y}(t)\mathrm{y}(t+\tau)} \\ &= \overline{\sum_{k=-\infty}^{\infty} \mathrm{a}_k p(t - kT_b - \alpha) \sum_{m=-\infty}^{\infty} \mathrm{a}_m p(t + \tau - mT_b - \alpha)} \\ &= \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \overline{\mathrm{a}_k \mathrm{a}_m p(t - kT_b - \alpha) p(t + \tau - mT_b - \alpha)} \end{aligned}$$

---

* If $\alpha = 0$, the process can be expressed as $\mathrm{y}(t) = \sum_{k=-\infty}^{\infty} \mathrm{a}_k p(t - kT_b)$. In this case $\overline{\mathrm{y}(t)} = \overline{\mathrm{a}_k} \sum_{k=-\infty}^{\infty} p(t - kT_b)$ is not constant, but is periodic with period T_b. Similarly, we can show that the autocorrelation function is periodic with the same period T_b. This is an example of a **cyclostationary**, or periodically stationary, process (a process whose statistics are invariant to a shift of the time origin by integral multiples of a constant T_b). Cyclostationary processes, as seen here, are clearly not wide-sense stationary. But they can be made wide-sense stationary with slight modification by adding the RV α in the expression of $\mathrm{y}(t)$, as in this example.

† Using exactly the same approach, as used in the derivation of Eq. (11.26) we can show that $\overline{\mathrm{y}(t)} = (\overline{\mathrm{a}_k}/T_b)\int_{-\infty}^{\infty} p(t)\,dt$.



(a)

(b)

Figure 11.10 Random PAM process.

Because a_k and a_m are independent of α,

$$R_y(\tau) = \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \overline{\mathrm{a}_k \mathrm{a}_m}\ \overline{p(t - kT_b - \alpha) p(t + \tau - mT_b - \alpha)}$$

Both k and m are integers. Letting $m = k + n$, this expression can be written

$$R_y(\tau) = \sum_{k=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \overline{\mathrm{a}_k \mathrm{a}_{k+n}}\ \overline{p(t - kT_b - \alpha) p(t + \tau - [k+n]T_b - \alpha)}$$

The first term under the double sum is the correlation of RVs a_k and a_{k+n} and will be denoted by $\mathfrak{R}_n$. The second term, being a mean with respect to the RV α, can be expressed as an integral. Thus,

$$R_y(\tau) = \sum_{n=-\infty}^{\infty} \mathfrak{R}_n \sum_{k=-\infty}^{\infty} \int_0^{T_b} p(t - kT_b - \alpha) p(t + \tau - [k+n]T_b - \alpha) p(\alpha)\, d\alpha$$

Recall that α is uniformly distributed over the interval 0 to T_b. Hence, $p(\alpha) = 1/T_b$ over the interval $(0, T_b)$, and is zero otherwise. Therefore,

$$\begin{aligned} R_y(\tau)7 &= \sum_{n=-\infty}^{\infty} \mathfrak{R}_n \sum_{k=-\infty}^{\infty} \frac{1}{T_b} \int_0^{T_b} p(t - kT_b - \alpha) p(t + \tau - [k+n]T_b - \alpha)\, d\alpha \\ &= \frac{1}{T_b} \sum_{n=-\infty}^{\infty} \mathfrak{R}_n \sum_{k=-\infty}^{\infty} \int_{t-(k+1)T_b}^{t-kT_b} p(\beta) p(\beta + \tau - nT_b)\, d\beta \\ &= \frac{1}{T_b} \sum_{n=-\infty}^{\infty} \mathfrak{R}_n \int_{-\infty}^{\infty} p(\beta) p(\beta + \tau - nT_b)\, d\beta \end{aligned}$$

The integral on the right-hand side is the time-autocorrelation function of the pulse $p(t)$ with the argument $\tau - nT_b$. Thus,

$$R_y(\tau) = \frac{1}{T_b} \sum_{n=-\infty}^{\infty} \mathfrak{R}_n \, \psi_p(\tau - nT_b) \tag{11.26}$$

where

$$\mathfrak{R}_n = \overline{a_k a_{k+n}} \tag{11.27}$$

and

$$\psi_p(\tau) = \int_{-\infty}^{\infty} p(t)p(t+\tau)\,dt \tag{11.28}$$

As seen in Eq. (3.73), if $p(t) \Longleftrightarrow P(\omega)$, then $\psi_p(\tau) \Longleftrightarrow |P(\omega)|^2$. Therefore, the PSD of $y(t)$, which is the Fourier transform of $R_y(\tau)$, is given by

$$\begin{aligned} S_y(\omega) &= \frac{1}{T_b} \sum_{n=-\infty}^{\infty} \mathfrak{R}_n |P(\omega)|^2 e^{-jn\omega T_b} \\ &= \frac{|P(\omega)|^2}{T_b} \sum_{n=-\infty}^{\infty} \mathfrak{R}_n e^{-jn\omega T_b} \end{aligned} \tag{11.29}$$

This result is similar to that found in Eq. (7.10b). The only difference is the use of the ensemble average in defining $\mathfrak{R}_n$ in this chapter, whereas R_n in Chapter 7 is the time average.

EXAMPLE 11.7 Find the PSD $S_y(\omega)$ for a polar binary random signal where **1** is transmitted by a pulse $p(t)$ (Fig. 11.11) whose Fourier transform is $P(\omega)$, and **0** is transmitted by $-p(t)$. The digits **1** and **0** are equally likely, and one digit is transmitted every T_b seconds. Each digit is independent of the other digits.

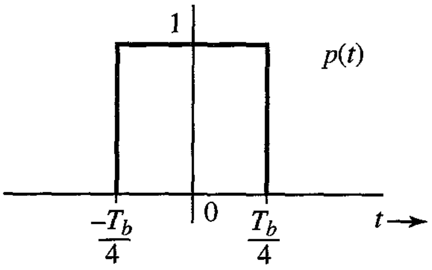


Figure 11.11 Basic pulse for a random binary polar signal.

In this case, a_k can take on values 1 and -1 with probability $1/2$ each. Hence,

$$\begin{aligned} \overline{a_k} &= \sum_{k=1,-1} a_k P(a_k) = (1)P_{a_k}(1) + (-1)P_{a_k}(-1) \\ &= \frac{1}{2} - \frac{1}{2} = 0 \end{aligned}$$

$$\begin{aligned} \mathfrak{R}_0 = \overline{a_k^2} &= \sum_{k=1,-1} a_k^2 P(a_k) = (1)^2 P_{a_k}(1) + (-1)^2 P_{a_k}(-1) \\ &= \frac{1}{2}(1)^2 + \frac{1}{2}(-1)^2 = 1 \end{aligned}$$

and because each digit is independent of the remaining digits,

$$\mathfrak{R}_n = \overline{a_k a_{k+n}} = \overline{a_k}\;\overline{a_{k+n}} = 0 \qquad n \geq 1$$

Hence, from Eq. (11.29),

$$S_y(\omega) = \frac{|P(\omega)|^2}{T_b}$$

We already found this result in Eq. (7.12), where we used time averaging instead of ensemble averaging. When a process is ergodic of second order (or higher), the ensemble and time averages yield the same result. Note that Example 11.5 is a special case of this result, where $p(t)$ is a full-width rectangular pulse $\text{rect}\,(t/T_b)$ with $P(\omega) = T_b\,\text{sinc}\,(\omega T_b/2)$, and

$$S_y(\omega) = \frac{|P(\omega)|^2}{T_b} = T_b\,\text{sinc}^2\left(\frac{\omega T_b}{2}\right)$$

EXAMPLE 11.8 Find the PSD $S_y(\omega)$ for on-off and bipolar random signals which use a basic pulse for $p(t)$, as shown in Fig. 11.11. The digits **1** and **0** are equally likely, and digits are transmitted every T_b seconds. Each digit is independent of the remaining digits. All these line codes are described in Sec. 7.2.

In each case we shall first determine $\mathfrak{R}_0$, $\mathfrak{R}_1$, $\mathfrak{R}_2$, $\cdots$, $\mathfrak{R}_n$.

(a) *On-off signaling:* In this case, a_n can take on values 1 and 0 with probability 1/2 each. Hence,

$$\overline{a_k} = (1)P_{a_k}(1) + (0)P_{a_k}(0) = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2}$$

$$\mathfrak{R}_0 = \overline{a_k^2} = (1)^2 P_{a_k}(1) + (0)^2 P_{a_k}(0) = \frac{1}{2}(1)^2 + \frac{1}{2}(0)^2 = \frac{1}{2}$$

and because each digit is independent of the remaining digits,

$$\mathfrak{R}_n = \overline{a_k a_{k+n}} = \overline{a_k}\;\overline{a_{k+n}} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4} \qquad n \geq 1$$

Therefore, from Eq. (11.29),

$$S_y(\omega) = \frac{|P(\omega)|^2}{T_b}\left[\frac{1}{2} + \frac{1}{4}\sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} e^{-jn\omega T_b}\right] \tag{11.30a}$$

$$= \frac{|P(\omega)|^2}{T_b}\left[\frac{1}{4} + \frac{1}{4}\sum_{n=-\infty}^{\infty} e^{-jn\omega T_b}\right] \tag{11.30b}$$

Equation (11.30b) is obtained from Eq. (11.30a) by splitting the term 1/2 corresponding to $\mathfrak{R}_0$ into two: 1/4 outside the summation and 1/4 inside the summation (corresponding to $n = 0$). This result is identical to Eq. (7.17b) found earlier by using time averages.

We now use the formula (see the footnote* for a proof),

$$\sum_{n=-\infty}^{\infty} e^{-jn\omega T_b} = \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)$$

Substitution of this result into Eq. (11.30b) yields

$$S_y(\omega) = \frac{|P(\omega)|^2}{4T_b}\left[1 + \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)\right] \tag{11.30c}$$

Note that the spectrum $S_y(\omega)$ consists of both a discrete and a continuous part. A discrete component of clock frequency ($R_b = 1/T_b$) is present in the spectrum. The continuous component of the spectrum is $|P(\omega)|^2/4T_b$ is identical (except for a scaling factor 1/4) to the spectrum of the polar signal in Example 11.7. This is a logical result because as Fig. 7.2 shows, an on-off signal can be expressed as a sum of a polar and a periodic component. The polar component is exactly half the polar signal discussed earlier. Hence, the PSD of this component is one-fourth of the PSD of the polar signal. The periodic component is of clock frequency R_b, and consists of discrete components of frequency R_b and its harmonics.

(b) *Bipolar signaling:* in this case, a_k can take on values 0, 1, and −1 with probabilities 1/2, 1/4, and 1/4, respectively. Hence,

$$\begin{aligned}\overline{\mathrm{a}_k} &= (0)P_{\mathrm{a}_k}(0) + (1)P_{\mathrm{a}_k}(1) + (-1)P_{\mathrm{a}_k}(-1)\\ &= \frac{1}{2}(0) + \frac{1}{4}(1) + \frac{1}{4}(-1) = 0\\ \mathfrak{R}_0 = \overline{\mathrm{a}_k^2} &= (0)^2P_{\mathrm{a}_k}(0) + (1)^2P_{\mathrm{a}_k}(1) + (-1)^2P_{\mathrm{a}_k}(-1)\\ &= \frac{1}{2}(0)^2 + \frac{1}{4}(1)^2 + \frac{1}{4}(-1)^2 = \frac{1}{2}\end{aligned}$$

Also,

$$\mathfrak{R}_1 = \overline{\mathrm{a}_k\mathrm{a}_{k+1}} = \sum_k \sum_{k+1} a_k a_{k+1} P_{\mathrm{a}_k\mathrm{a}_{k+1}}(a_k a_{k+1})$$

Because a_k and a_{k+1} can take three values each, the sum on the right-hand side has nine terms, of which only four terms (corresponding to values ±1 for a_k and a_{k+1}) are nonzero. Thus,

$$\begin{aligned}\mathfrak{R}_1 = {}&(1)(1)P_{\mathrm{a}_k\mathrm{a}_{k+1}}(1,\ 1) + (-1)(1)P_{\mathrm{a}_k\mathrm{a}_{k+1}}(-1,\ 1)\\ &+ (1)(-1)P_{\mathrm{a}_k\mathrm{a}_{k+1}}(1,\ -1) + (-1)(-1)P_{\mathrm{a}_k\mathrm{a}_{k+1}}(-1,\ -1)\end{aligned}$$

---

[*] The impulse train in Fig. 3.24a is $\delta_{T_b}(t)$ can be expressed as $\delta_{T_b}(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_b)$. Also $\delta(t - nT_b) \Leftrightarrow e^{-jn\omega T_b}$. Hence, the Fourier transform of this impulse train is $\sum_{n=-\infty}^{\infty} e^{-jn\omega T_b}$. But we found the alternate form of the Fourier transform of this train in Eq. (3.42) (Example 3.13). Hence,

$$\sum_{n=-\infty}^{\infty} e^{jn\omega T_b} = \frac{2\pi}{T_b} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi n}{T_b}\right)$$

Because of the bipolar rule,

$$P_{a_k a_{k+1}}(1,\ 1) = P_{a_k a_{k+1}}(-1,\ -1) = 0$$

and

$$P_{a_k a_{k+1}}(-1,\ 1) = P_{a_k}(-1) P_{a_{k+1}|a_k}(1|-1) = \left(\frac{1}{4}\right)\left(\frac{1}{2}\right) = \frac{1}{8}$$

Similarly, we find $P_{a_k a_{k+1}}(1,\ -1) = 1/8$. Substitution of these values in $\mathfrak{R}_1$ yields

$$\mathfrak{R}_1 = -\frac{1}{4}$$

For $n \geq 2$, the pulse strengths a_k and a_{k+1} become independent. Hence,

$$\mathfrak{R}_n = \overline{a_k a_{k+n}} = \overline{a_k}\ \overline{a_{k+n}} = (0)(0) = 0 \qquad n \geq 2$$

Substitution of these values in Eq. (11.29) and noting that $\mathfrak{R}_n$ is an even function of n, yields

$$S_y(\omega) = \frac{|P(\omega)|^2}{T_b} \sin^2\left(\frac{\omega T_b}{2}\right)$$

This result is identical to Eq. (7.20b) found earlier by using time averages.

11.3 MULTIPLE RANDOM PROCESSES

For two real random processes $x(t)$ and $y(t)$, we define the **crosscorrelation function*** $R_{xy}(t_1,\ t_2)$ as

$$R_{xy}(t_1,\ t_2) = \overline{x(t_1)y(t_2)} \tag{11.31a}$$

The two processes are said to be **jointly stationary** (in the wide sense) if each of the processes is individually wide-sense stationary and if

$$\begin{aligned} R_{xy}(t_1,\ t_2) &= R_{xy}(t_2 - t_1) \\ &= R_{xy}(\tau) \end{aligned} \tag{11.31b}$$

Uncorrelated, Orthogonal (Incoherent), and Independent Processes

Two processes $x(t)$ and $y(t)$ are said to be **uncorrelated** if their crosscorrelation function is equal to the product of their means; that is,

$$R_{xy}(\tau) = \overline{x(t)y(t+\tau)} = \bar{x}\bar{y} \tag{11.32}$$

This implies that RVs $x(t)$ and $y(t+\tau)$ are uncorrelated for all t and τ.

---

* For complex random processes, the crosscorrelation function is defined as

$$R_{xy}(t_1,\ t_2) = \overline{x^*(t_1)y(t_2)} \qquad (11.31a)$$

Processes $x(t)$ and $y(t)$ are said to be **incoherent**, or **orthogonal**, if

$$R_{xy}(\tau) = 0 \tag{11.33}$$

Incoherent, or orthogonal, processes are uncorrelated processes with $\bar{x}$ and/or $\bar{y} = 0$.

Processes $x(t)$ and $y(t)$ are **independent** random processes if for any t_1 and t_2, $x(t_1)$ and $y(t_2)$ are independent.

Cross-Power Spectral Density

We define the **cross-power spectral density** $S_{xy}(\omega)$ for two random processes $x(t)$ and $y(t)$ as

$$S_{xy}(\omega) = \lim_{T \to \infty} \frac{\overline{X_T^*(\omega) Y_T(\omega)}}{T} \tag{11.34}$$

where $X_T(\omega)$ and $Y_T(\omega)$ are the Fourier transforms of the truncated processes $x(t)\,\text{rect}\,(t/T)$ and $y(t)\,\text{rect}\,(t/T)$, respectively. Proceeding along the lines of the derivation of Eq. (11.14), it can be shown that*

$$R_{xy}(\tau) \Longleftrightarrow S_{xy}(\omega) \tag{11.35a}$$

It can be seen from Eqs. (11.31) that for real random processes $x(t)$ and $y(t)$,

$$R_{xy}(\tau) = R_{yx}(-\tau) \tag{11.35b}$$

Therefore,

$$S_{xy}(\omega) = S_{yx}(-\omega) \tag{11.35c}$$

11.4 TRANSMISSION OF RANDOM PROCESSES THROUGH LINEAR SYSTEMS

If a random process $x(t)$ is applied at the input of a linear time-invariant system (Fig. 11.12) with transfer function $H(\omega)$, we can determine the autocorrelation function and the PSD of the output process $y(t)$. We now show that

$$R_y(\tau) = h(\tau) * h(-\tau) * R_x(\tau) \tag{11.36}$$

and

$$S_y(\omega) = |H(\omega)|^2 \, S_x(\omega) \tag{11.37}$$

To prove this, we observe that

$$y(t) = \int_{-\infty}^{\infty} h(\alpha) x(t - \alpha)\, d\alpha$$

and

$$y(t + \tau) = \int_{-\infty}^{\infty} h(\alpha) x(t + \tau - \alpha)\, d\alpha$$

---

* Equation (11.35a) is valid for complex processes as well.

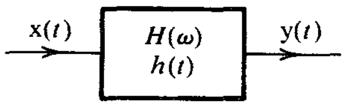


Figure 11.12 Transmission of a random process through a linear time-invariant system.

Hence,*

$$R_y(\tau) = \overline{y(t)y(t+\tau)} = \overline{\int_{-\infty}^{\infty} h(\alpha)x(t-\alpha)\,d\alpha \int_{-\infty}^{\infty} h(\beta)x(t+\tau-\beta)\,d\beta}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)\overline{x(t-\alpha)x(t+\tau-\beta)}\,d\alpha\,d\beta$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)R_x(\tau+\alpha-\beta)\,d\alpha\,d\beta$$

This double integral is precisely the double convolution $h(\tau) * h(-\tau) * R_x(\tau)$. Hence, Eqs. (11.36) and (11.37) follow.

EXAMPLE 11.9 Thermal Noise

Random thermal motion of electrons in a resistor R causes a random voltage across its terminals. This voltage $n(t)$ is known as the **thermal noise**. Its PSD $S_n(\omega)$ is practically flat over a very large band (up to 1000 GHz at room temperature) and is given by[1]

$$S_n(\omega) = 2kTR \tag{11.38}$$

where k is the Boltzmann constant (1.38×10^{-23}) and T is the ambient temperature in kelvins. A resistor R at a temperature TK can be represented by a noiseless resistor R in series with a random white-noise voltage source (thermal noise) having a PSD of $2kTR$ (Fig. 11.13a). Observe that the thermal noise power over a band Δf is $(2kTR)\,2\Delta f = 4kTR\Delta f$.

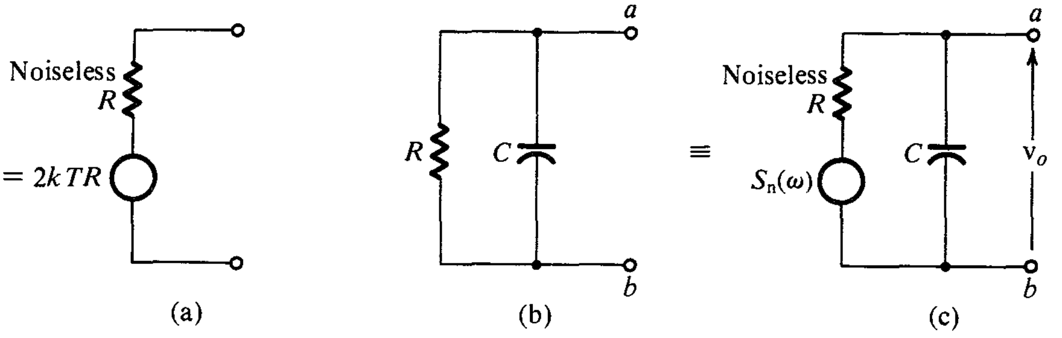


Figure 11.13 Thermal noise representation in a resistor.

Let us calculate the thermal noise voltage (rms value) across the simple RC circuit in Fig. 11.13b. The resistor R is replaced by an equivalent noiseless resistor in series with the thermal-noise voltage source. The transfer function $H(\omega)$ relating the voltage v_o at terminals ab to the thermal noise voltage is given by

---

* In this development, we interchange the operations of averaging and integrating. Because averaging is really an operation of integration, we are really changing the order of integration, and we assume that such a change is permissible.

$$H(\omega) = \frac{1/j\omega C}{R + 1/j\omega C} = \frac{1}{1 + j\omega RC}$$

If $S_0(\omega)$ is the PSD of the voltage v_o, then from Eq. (11.37) we have

$$\begin{aligned} S_0(\omega) &= \left|\frac{1}{1 + j\omega RC}\right|^2 2kTR \\ &= \frac{2kTR}{1 + \omega^2 R^2 C^2} \end{aligned}$$

The mean square value $\overline{\mathrm{v}_o^2}$ is given by

$$\begin{aligned} \overline{\mathrm{v}_o^2} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{2kTR}{1 + \omega^2 R^2 C^2}\, d\omega \\ &= \frac{kT}{C} \end{aligned} \tag{11.39}$$

Hence, the rms thermal noise voltage across the capacitor is $\sqrt{kT/C}$.

Sum of Random Processes

If two stationary processes (at least in the wide sense) $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are added to form a process $\mathrm{z}(t)$, the statistics of $\mathrm{z}(t)$ can be determined in terms of those of $\mathrm{x}(t)$ and $\mathrm{y}(t)$. If

$$\mathrm{z}(t) = \mathrm{x}(t) + \mathrm{y}(t) \tag{11.40a}$$

then

$$\begin{aligned} R_{\mathrm{z}}(\tau) &= \overline{\mathrm{z}(t)\mathrm{z}(t+\tau)} = \overline{[\mathrm{x}(t) + \mathrm{y}(t)][\mathrm{x}(t+\tau) + \mathrm{y}(t+\tau)]} \\ &= R_{\mathrm{x}}(\tau) + R_{\mathrm{y}}(\tau) + R_{\mathrm{xy}}(\tau) + R_{\mathrm{yx}}(\tau) \end{aligned} \tag{11.40b}$$

If $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are uncorrelated, then from Eq. (11.32),

$$R_{\mathrm{xy}}(\tau) = R_{\mathrm{yx}}(\tau) = \bar{\mathrm{x}}\bar{\mathrm{y}}$$

and

$$R_{\mathrm{z}}(\tau) = R_{\mathrm{x}}(\tau) + R_{\mathrm{y}}(\tau) + 2\bar{\mathrm{x}}\bar{\mathrm{y}} \tag{11.41}$$

Most processes of interest in communication problems have zero means. If processes $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are uncorrelated with either $\bar{\mathrm{x}}$ or $\bar{\mathrm{y}} = 0$ (that is, if $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are incoherent), then

$$R_{\mathrm{z}}(\tau) = R_{\mathrm{x}}(\tau) + R_{\mathrm{y}}(\tau) \tag{11.42a}$$

and

$$S_{\mathrm{z}}(\omega) = S_{\mathrm{x}}(\omega) + S_{\mathrm{y}}(\omega) \tag{11.42b}$$

It also follows from Eqs. (11.42a) and (11.17) that

$$\overline{\mathrm{z}^2} = \overline{\mathrm{x}^2} + \overline{\mathrm{y}^2} \tag{11.42c}$$

Hence, the mean square of a sum of incoherent (or orthogonal) processes is equal to the sum of the mean squares of these processes.

EXAMPLE 11.10 Two independent random voltage processes $x_1(t)$ and $x_2(t)$ are applied to an *RC* network, as shown in Fig. 11.14. It is given that

$$S_{x_1}(\omega) = K \qquad S_{x_2}(\omega) = \frac{2\alpha}{\alpha^2 + \omega^2}$$

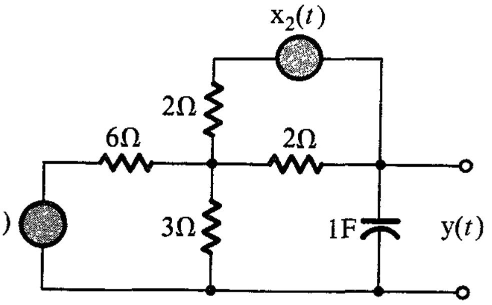


Figure 11.14 Noise calculations in a resistive circuit.

Determine the PSD and the power P_y of the output random process $y(t)$. Assume that the resistors in the circuit contribute negligible thermal noise (i.e., assume that they are noiseless).

Because the network is linear, the output voltage $y(t)$ can be expressed as

$$y(t) = y_1(t) + y_2(t)$$

where $y_1(t)$ is the output from input $x_1(t)$ [assuming $x_2(t) = 0$] and $y_2(t)$ is the output from input $x_2(t)$ [assuming $x_1(t) = 0$]. The transfer functions relating $y(t)$ to $x_1(t)$ and $x_2(t)$ are $H_1(\omega)$ and $H_2(\omega)$, respectively, given by

$$H_1(\omega) = \frac{1}{3(3j\omega + 1)}, \qquad H_2(\omega) = \frac{1}{2(3j\omega + 1)}$$

Hence,

$$S_{y_1}(\omega) = |H_1(\omega)|^2 S_{x_1}(\omega) = \frac{K}{9(9\omega^2 + 1)}$$

and

$$S_{y_2}(\omega) = |H_2(\omega)|^2 S_{x_2}(\omega) = \frac{\alpha}{2(9\omega^2 + 1)(\alpha^2 + \omega^2)}$$

Because the input processes $x_1(t)$ and $x_2(t)$ are independent, the outputs $y_1(t)$ and $y_2(t)$ generated by them will also be independent. Also, the PSDs of $y_1(t)$ and $y_2(t)$ have no impulses at $\omega = 0$, implying that they have no dc components [i.e., $\overline{y_1(t)} = \overline{y_2(t)} = 0$]. Hence, $y_1(t)$ and $y_2(t)$ are incoherent, and

$$\begin{aligned} S_y(\omega) &= S_{y_1}(\omega) + S_{y_2}(\omega) \\ &= \frac{2K(\alpha^2 + \omega^2) + 9\alpha}{18(9\omega^2 + 1)(\alpha^2 + \omega^2)} \end{aligned}$$

The power P_y (or the mean square value $\overline{y^2}$) can be determined in two ways. We can find $R_y(\tau)$ by taking the inverse transforms of $S_{y_1}(\omega)$ and $S_{y_2}(\omega)$ as

$$R_y(\tau) = \underbrace{\frac{K}{54}e^{-|\tau|/3}}_{R_{y_1}(\tau)} + \underbrace{\frac{3\alpha - e^{-\alpha|\tau|}}{4(9\alpha^2 - 1)}}_{R_{y_2}(\tau)}$$

and

$$P_y = \overline{y^2} = R_y(0) = \frac{K}{54} + \frac{3\alpha - 1}{4(9\alpha^2 - 1)}$$

Alternatively, we can determine $\overline{y^2}$ by integrating $S_y(\omega)$ with respect to ω (or f) [see Eq. (11.17)].

11.5 BANDPASS RANDOM PROCESSES

If the PSD of a random process is confined to a certain passband (Fig. 11.15), the process is a **bandpass** random process. Just as a bandpass signal can be represented in terms of quadrature components [see Eq. (3.38)], we can express a bandpass random process $x(t)$ in terms of quadrature components as follows:

$$x(t) = x_c(t)\cos\omega_c t + x_s(t)\sin\omega_c t \tag{11.43}$$

This can be proved by considering the system in Fig. 11.16a, where $H_0(\omega)$ is an ideal low-pass filter (Fig. 11.16b) with unit impulse response $h_0(t)$. First we show that the system in Fig. 11.16a is an ideal bandpass filter with the transfer function $H(\omega)$ shown in Fig. 11.16c. This can be conveniently done by computing the response $h(t)$ to the unit impulse input $\delta(t)$. Because the system contains time-varying multipliers, however, we must also test whether it is a time-varying or a time-invariant system. It is therefore appropriate to consider the system response to an input $\delta(t-\alpha)$. This is an impulse at $t = \alpha$. Using the fact that [see Eq. (2.18b)] $f(t)\,\delta(t-\alpha) = f(\alpha)\delta(t-\alpha)$, we can express the signals at various points as follows:

Signal at a_1	$2\cos(\omega_c\alpha + \theta)\,\delta(t-\alpha)$
a_2	$2\sin(\omega_c\alpha + \theta)\,\delta(t-\alpha)$
b_1	$2\cos(\omega_c\alpha + \theta)h_0(t-\alpha)$
b_2	$2\sin(\omega_c\alpha + \theta)h_0(t-\alpha)$
c_1	$2\cos(\omega_c\alpha + \theta)\cos(\omega_c t + \theta)h_0(t-\alpha)$
c_2	$2\sin(\omega_c\alpha + \theta)\sin(\omega_c t + \theta)h_0(t-\alpha)$
d	$2h_0(t-\alpha)\,[\cos(\omega_c\alpha + \theta)\cos(\omega_c t + \theta) + \sin(\omega_c\alpha + \theta)\sin(\omega_c t + \theta)]$ $= 2h_0(t-\alpha)\cos[\omega_c(t-\alpha)]$

Thus, the system response to the input $\delta(t-\alpha)$ is $2h_0(t-\alpha)\cos[\omega_c(t-\alpha)]$. Clearly, the system is linear time-invariant, with impulse response

$$h(t) = 2h_0(t)\cos\omega_c t$$

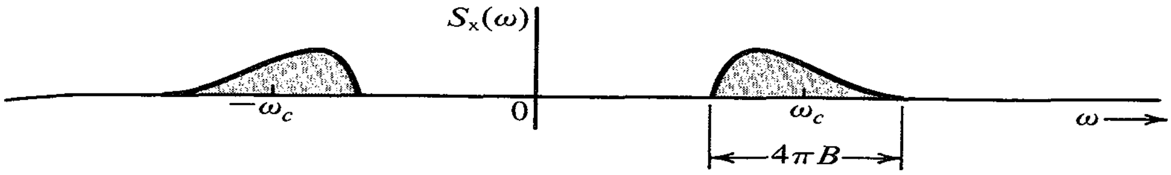


Figure 11.15 PSD of a bandpass random process.

and

$$H(\omega) = H_0(\omega + \omega_c) + H_0(\omega - \omega_c)$$

The transfer function $H(\omega)$ (Fig. 11.16c) represents an ideal bandpass filter.

If we apply the bandpass process $x(t)$ (Fig. 11.15) to the input of this system, the output will be $x(t)$. If the processes at points b_1 and b_2 (low-pass filter outputs) are denoted by $x_c(t)$ and $x_s(t)$, respectively, then the output $x(t)$ can be written as

$$x(t) = x_c(t)\cos(\omega_c t + \theta) + x_s(t)\sin(\omega_c t + \theta) \tag{11.44}$$

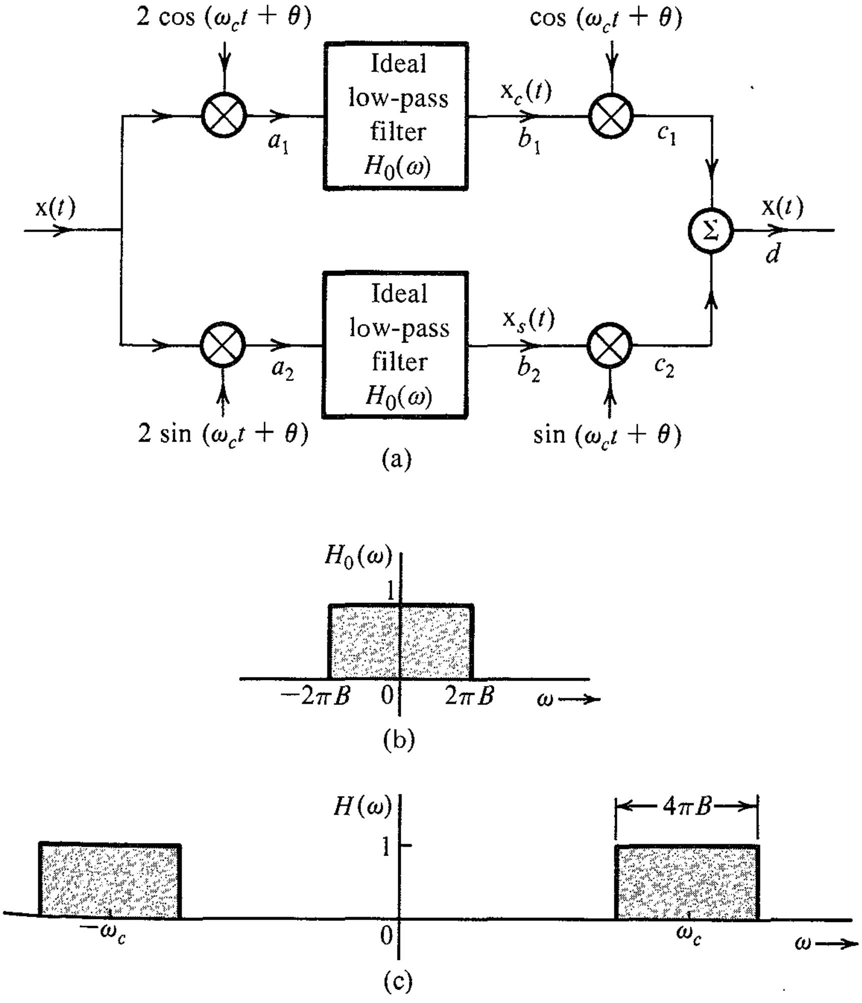

$2\cos(\omega_c t + \theta)$
$\cos(\omega_c t + \theta)$
Ideal low-pass filter $H_0(\omega)$
$x_c(t)$
a_1
b_1
c_1
$x(t)$
Σ
$x(t)$
d
Ideal low-pass filter $H_0(\omega)$
$x_s(t)$
a_2
b_2
c_2
$2\sin(\omega_c t + \theta)$
$\sin(\omega_c t + \theta)$
(a)
$H_0(\omega)$
1
$-2\pi B$
0
$2\pi B$
ω
(b)
$H(\omega)$
1
$4\pi B$
$-\omega_c$
0
ω_c
ω
(c)

Figure 11.16 (a) Equivalent circuit of an ideal bandpass filter. (b) Ideal low-pass filter frequency response. (c) Ideal bandpass filter frequency response.

where $\mathrm{x}_c(t)$ and $\mathrm{x}_s(t)$ are low-pass random processes band-limited to B Hz (because they are the outputs of low-pass filters of bandwidth B). Because Eq. (11.44) is valid for any value of θ, by substituting $\theta = 0$, we get the desired representation in Eq. (11.43).

In order to characterize $\mathrm{x}_c(t)$ and $\mathrm{x}_s(t)$, consider once again Fig. 11.16a with the input $\mathrm{x}(t)$. Let θ be an RV uniformly distributed over the range $(0,\ 2\pi)$, that is, for a sample function, θ is equally likely to take on any value in the range $(0,\ 2\pi)$. In this case $\mathrm{x}(t)$ is represented as in Eq. (11.44). We observe that $\mathrm{x}_c(t)$ is obtained by multiplying $\mathrm{x}(t)$ by $2\cos\,(\omega_c t + \theta)$, and then passing the result through a low-pass filter. The PSD of $2\mathrm{x}(t)\cos\,(\omega_c t + \theta)$ is [see Eq. (11.20b)]

$$4 \times \frac{1}{4}[S_\mathrm{x}(\omega + \omega_c) + S_\mathrm{x}(\omega - \omega_c)]$$

This PSD is $S_\mathrm{x}(\omega)$ shifted up and down by ω_c, as shown in Fig. 11.17a. When this is passed through a low-pass filter, the resulting PSD of $\mathrm{x}_c(t)$ is as shown in Fig. 11.17b. It is clear that

$$S_{\mathrm{x}_c}(\omega) = \begin{cases} S_\mathrm{x}(\omega + \omega_c) + S_\mathrm{x}(\omega - \omega_c) & |\omega| \le 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \tag{11.45a}$$

We can obtain $S_{\mathrm{x}_s}(\omega)$ in the same way. As far as the PSD is concerned, multiplication by $\cos\,(\omega_c t + \theta)$ or $\sin\,(\omega_c t + \theta)$ makes no difference (see footnote following Eq. (11.20a), and we get

$$S_{\mathrm{x}_c}(\omega) = S_{\mathrm{x}_s}(\omega) = \begin{cases} S_\mathrm{x}(\omega + \omega_c) + S_\mathrm{x}(\omega - \omega_c) & |\omega| \le 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \tag{11.45b}$$

From Figs. 11.15 and 11.17b, we make the interesting observation that the areas under the PSDs $S_\mathrm{x}(\omega)$, $S_{\mathrm{x}_c}(\omega)$, and $S_{\mathrm{x}_s}(\omega)$ are equal. Hence, it follows that

$$\overline{\mathrm{x}_c^2(t)} = \overline{\mathrm{x}_s^2(t)} = \overline{\mathrm{x}^2(t)} \tag{11.45c}$$

Thus, the mean square values (or powers) of $\mathrm{x}_c(t)$ and $\mathrm{x}_s(t)$ are identical to that of $\mathrm{x}(t)$.

These results are derived by assuming Θ to be an RV. For the representation in Eq. (11.43), $\Theta = 0$, and Eqs. (11.45b, c) may not be true. Fortunately, Eqs. (11.45b, c) hold even for the

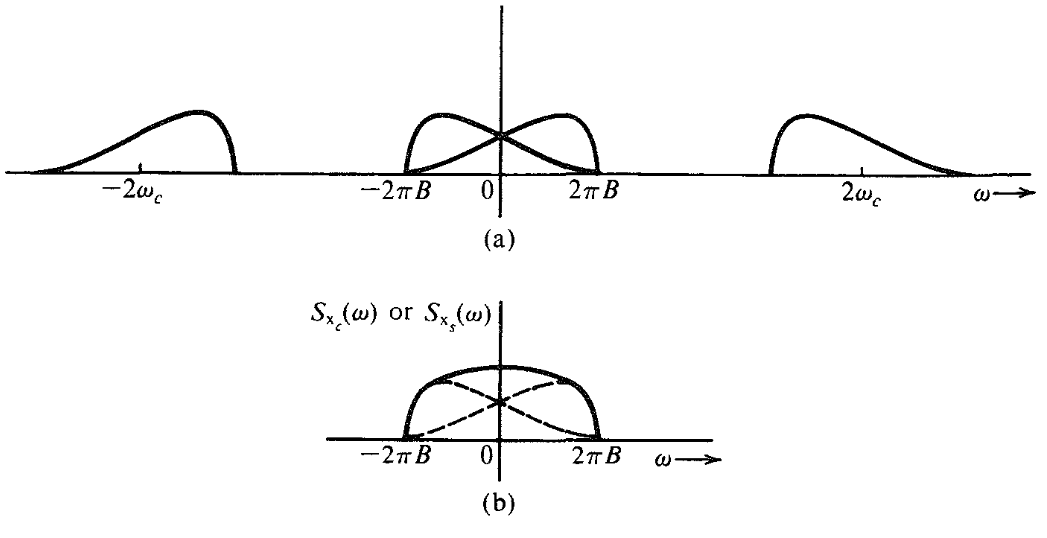


Figure 11.17 Derivation of PSDs of quadrature components of a bandpass random process.

case of $\Theta = 0$. The proof is rather long and cumbersome and will not be given here.[1,2,3] It can also be shown that[1,2,3]

$$\overline{\mathrm{x}_c(t)\mathrm{x}_s(t)} = R_{\mathrm{x}_c\mathrm{x}_s}(0) = 0 \tag{11.46}$$

That is, the amplitudes x_c and x_s at any given instant are uncorrelated. Moreover, if $S_\mathrm{x}(\omega)$ is symmetrical about ω_c (as well as $-\omega_c$), then

$$R_{\mathrm{x}_c\mathrm{x}_s}(\tau) = 0 \tag{11.47}$$

EXAMPLE 11.11 The PSD of a bandpass white noise $\mathrm{n}(t)$ is $\mathcal{N}/2$ (Fig. 11.18a). Represent this process in terms of quadrature components. Derive $S_{\mathrm{n}_c}(\omega)$ and $S_{\mathrm{n}_s}(\omega)$, and verify that $\overline{\mathrm{n}_c^2} = \overline{\mathrm{n}_s^2} = \overline{\mathrm{n}^2}$.

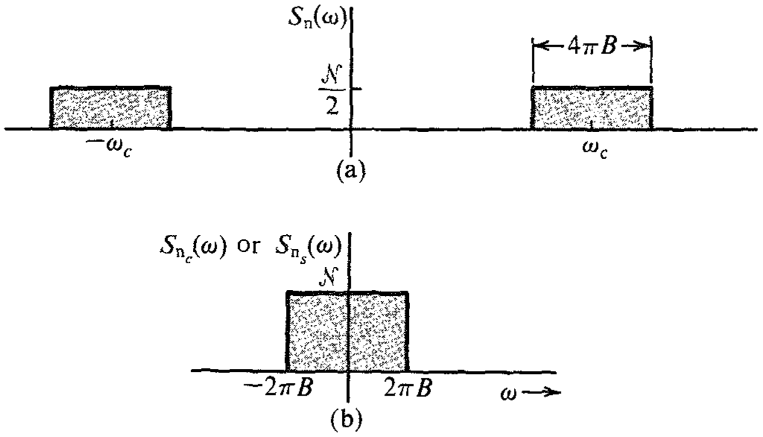


Figure 11.18 (a) PSD of a bandpass white-noise process. (b) PSD of its quadrature components.

We have the expression

$$\mathrm{n}(t) = \mathrm{n}_c(t)\cos\omega_c t + \mathrm{n}_s(t)\sin\omega_c t \tag{11.48}$$

where

$$S_{\mathrm{n}_c}(\omega) = S_{\mathrm{n}_s}(\omega) = \begin{cases} S_\mathrm{n}(\omega+\omega_c) + S_\mathrm{n}(\omega-\omega_c) & |\omega| \le 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases}$$

It follows from this equation and from Fig. 11.18 that

$$S_{\mathrm{n}_c}(\omega) = S_{\mathrm{n}_s}(\omega) = \begin{cases} \mathcal{N} & |\omega| \le 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \tag{11.49}$$

Also,

$$\overline{\mathrm{n}^2} = 2\int_{f_c-B}^{f_c+B} \frac{\mathcal{N}}{2}\,df = 2\mathcal{N}B \tag{11.50a}$$

From Fig. 11.18b it follows that

$$\overline{\mathrm{n}_c^2} = \overline{\mathrm{n}_s^2} = 2\int_0^B \mathcal{N}\,df = 2\mathcal{N}B \tag{11.50b}$$

Hence,

$$\overline{n_c^2} = \overline{n_s^2} = \overline{n^2} = 2\mathcal{N}B \tag{11.50c}$$

Nonuniqueness of the Quadrature Representation

No unique center frequency exists for a bandpass signal. For the spectrum in Fig. 11.19a, for example, we may consider the spectrum to have a bandwidth $2B$ centered at ω_c. The same spectrum can be considered to have a bandwidth $2B'$ centered at ω_1, as also shown in Fig. 11.19a. The quadrature representation [Eq. (11.43)] is also possible for center frequency ω_1:

$$x(t) = x_{c_1}(t)\cos\omega_1 t + x_{s_1}(t)\,\sin\omega_1(t)$$

where

$$S_{x_{c_1}}(\omega) = S_{x_{s_1}}(\omega) = \begin{cases} S_x(\omega+\omega_1) + S_x(\omega-\omega_1) & |\omega| \le 2\pi B' \\ 0 & |\omega| > 2\pi B' \end{cases} \tag{11.51}$$

This is shown in Fig. 11.19b. Thus, the quadrature representation of a bandpass process is not unique. An infinite number of possible choices exist for the center frequency, and corresponding to each center frequency is a distinct quadrature representation.

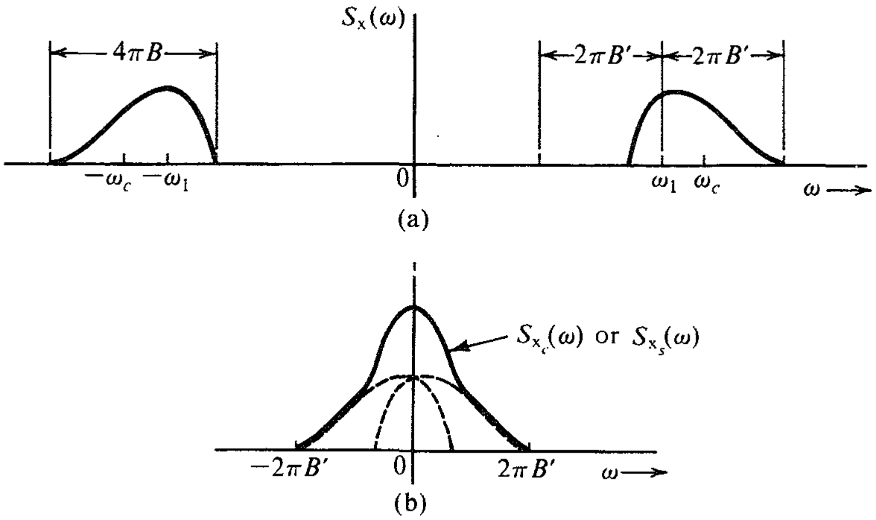


Figure 11.19 Nonunique nature of quadrature component representation of a bandpass process.

EXAMPLE 11.12 A bandpass white-noise PSD of an SSB channel (lower sideband) is shown in Fig. 11.20a. Represent this signal in terms of quadrature components with the carrier frequency ω_c.

The true center frequency of this PSD is not ω_c, but we can still use ω_c as the center frequency, as discussed earlier,

$$n(t) = n_c(t)\cos\omega_c t + n_s(t)\sin\omega_c t \tag{11.52}$$

The PSD $S_{n_c}(\omega)$ or $S_{n_s}(\omega)$ obtained by shifting $S_n(\omega)$ up and down by ω_c [see Eq. 11.51)] is shown in Fig. 11.20b,

$$S_{n_c}(\omega) = S_{n_s}(\omega) = \begin{cases} \dfrac{\mathcal{N}}{2} & |\omega| \le 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \tag{11.53}$$

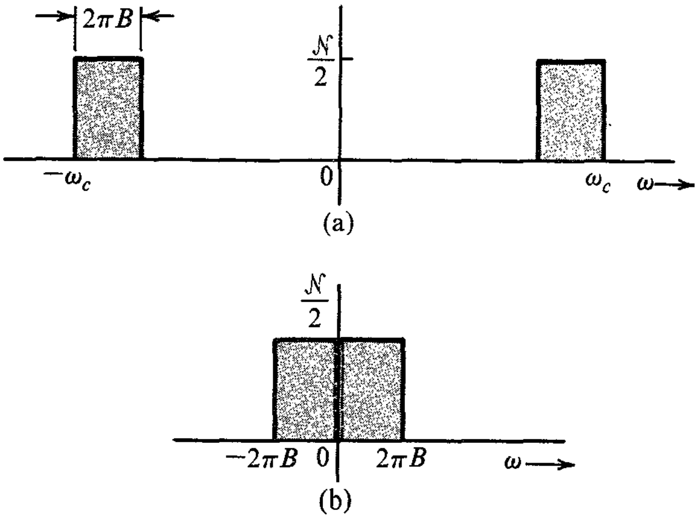


Figure 11.20 A possible form of quadrature component representation of noise in SSB.

From Fig. 11.20a it follows that

$$\overline{n^2} = \mathcal{N}B \tag{11.54a}$$

Similarly, from Fig. 11.20b we have

$$\overline{n_c^2} = \overline{n_s^2} = \mathcal{N}B \tag{11.54b}$$

Hence,

$$\overline{n_c^2} = \overline{n_s^2} = \overline{n^2} = \mathcal{N}B \tag{11.54c}$$

Bandpass White Gaussian Random Process

Thus far we have avoided defining a gaussian random process. The gaussian random process is perhaps the single most important random process in the area of communication. It requires a rather careful and unhurried discussion. Fortunately, we do not need to know much about the gaussian process at this point; therefore, its detailed discussion is postponed until Chapter 14 in order to avoid unnecessary digression. All we need to know here is that an RV $x(t)$ formed by sample-function amplitudes at instant t of a gaussian process is gaussian, with a PDF of the form of Eq. (10.38).

A gaussian random process with a uniform PSD is called a white gaussian random process. A bandpass white gaussian process $n(t)$ with PSD $\mathcal{N}/2$ centered at ω_c and with a bandwidth $2B$ (Fig. 11.18a) can be expressed in terms of quadrature components as

$$n(t) = n_c(t)\cos\omega_c t + n_s(t)\sin\omega_c t \tag{11.55}$$

where, from Eq. (11.49), we have

$$S_{n_c}(\omega) = S_{n_s}(\omega) = \begin{cases} \mathcal{N} & |\omega| \leq 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases}$$

Also, from Eq. (11.50c),

$$\overline{n_c^2} = \overline{n_s^2} = \overline{n^2} = 2\mathcal{N}B \tag{11.56}$$

The bandpass signal can also be expressed in polar form [see Eq. (3.39)]:

$$n(t) = E(t) \cos (\omega_c t + \Theta) \tag{11.57a}$$

where

$$E(t) = \sqrt{n_c^2(t) + n_s^2(t)} \tag{11.57b}$$

$$\Theta(t) = -\tan^{-1} \frac{n_s(t)}{n_c(t)} \tag{11.57c}$$

The RVs $n_c(t)$ and $n_s(t)$ are uncorrelated [see Eq. (11.46)] gaussian RVs with zero means and variance $2\mathcal{N}B$ [Eq. (11.56)]. Hence, their PDFs are identical:

$$p_{n_c}(\alpha) = p_{n_s}(\alpha) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\alpha^2/2\sigma^2} \tag{11.58a}$$

where

$$\sigma^2 = 2\mathcal{N}B \tag{11.58b}$$

It will also be shown in Chapter 14 that if two gaussian RVs are uncorrelated, they are independent. In such a case, as shown in Example 10.14, $E(t)$ has a Rayleigh density

$$p_E(E) = \frac{E}{\sigma^2} e^{-E^2/2\sigma^2} \qquad \sigma^2 = 2\mathcal{N}B \tag{11.59}$$

and Θ in Eq. (11.57a) is uniformly distributed over $(0,\ 2\pi)$.

Sinusoidal Signal in Noise

Another case of interest is a sinusoid plus a narrow-band gaussian noise. If $A \cos (\omega_c t + \varphi)$ is a sinusoid mixed with $n(t)$, a gaussian bandpass noise centered at ω_c, then the sum $y(t)$ is given by

$$y(t) = A \cos (\omega_c t + \varphi) + n(t)$$

Using Eq. (11.44) to represent the bandpass noise, we have

$$y(t) = [A + n_c(t)] \cos (\omega_c t + \varphi) + n_s(t) \sin (\omega_c t + \varphi) \tag{11.60a}$$

$$= E(t) \cos [\omega_c t + \Theta(t) + \varphi] \tag{11.60b}$$

where $E(t)$ is the envelope $[E(t) > 0]$ and $\Theta(t)$ is the angle shown in Fig. 11.21,

$$E(t) = \sqrt{[A + n_c(t)]^2 + n_s^2(t)} \tag{11.61a}$$

$$\Theta(t) = -\tan^{-1} \frac{n_s(t)}{A + n_c(t)} \tag{11.61b}$$

Both $n_c(t)$ and $n_s(t)$ are gaussian, with variance σ^2. For white gaussian noise, $\sigma^2 = 2\mathcal{N}B$ [Eq. (11.58b)]. Arguing in a manner analogous to that used in deriving Eq. (10.53), and observing that

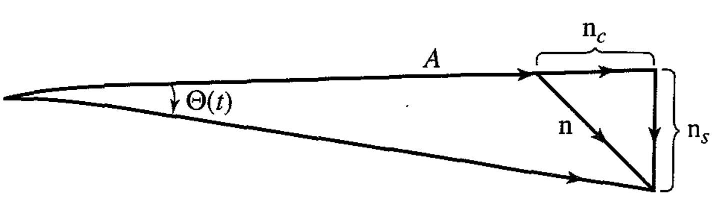


Figure 11.21 Phasor representation of a sinusoid and a narrow-band gaussian noise.

$$\begin{aligned}\mathrm{n}_c^2 + \mathrm{n}_s^2 &= E^2 - A^2 - 2A\mathrm{n}_c \\ &= E^2 - 2A(A + \mathrm{n}_c) + A^2 \\ &= E^2 - 2A\mathrm{E} \cos \Theta(t) + A^2\end{aligned}$$

we have

$$p_{\mathrm{E}\Theta}(E,\ \theta) = \frac{E}{2\pi\sigma^2} e^{-(E^2 - 2AE\cos\theta + A^2)/2\sigma^2} \tag{11.62}$$

where σ^2 is the variance of n_c (or n_s) and is equal to $2\mathcal{N}B$ for white noise. From Eq. (11.62) we have

$$\begin{aligned}p_{\mathrm{E}}(E) &= \int_{-\pi}^{\pi} p_{\mathrm{E}\Theta}(E, \theta)\, d\theta \\ &= \frac{E}{\sigma^2} e^{-(E^2 + A^2)/2\sigma^2} \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{(AE/\sigma^2)\cos\theta}\, d\theta\right]\end{aligned} \tag{11.63}$$

The bracketed term on the right-hand side of Eq. (11.63) defines $I_0(AE/\sigma^2)$, where I_0 is the **modified zero-order Bessel function** of the first kind. Thus,

$$p_{\mathrm{E}}(E) = \frac{E}{\sigma^2} e^{-(E^2 + A^2)/2\sigma^2} I_0\left(\frac{AE}{\sigma^2}\right) \tag{11.64a}$$

This is known as the **Rice density**, or **rician density**. For a large sinusoidal signal ($A \gg \sigma$), it can be shown that[4]

$$I_0\left(\frac{AE}{\sigma^2}\right) \simeq \sqrt{\frac{\sigma^2}{2\pi AE}}\, e^{AE/\sigma^2}$$

and

$$P_{\mathrm{E}}(E) \simeq \sqrt{\frac{E}{2\pi A\sigma^2}}\, e^{-(E-A)^2/2\sigma^2} \tag{11.64b}$$

Because $A \gg \sigma$, $E \simeq A$, and $p_{\mathrm{E}}(E)$ in Eq. (11.64b) is very nearly a gaussian density with mean A and variance σ,

$$p_{\mathrm{E}}(E) \simeq \frac{1}{\sigma\sqrt{2\pi}} e^{-(E-A)^2/2\sigma^2} \tag{11.64c}$$

Figure 11.22 shows the PDF of the normalized RV E/σ. Note that for $A/\sigma = 0$, we obtain the Rayleigh density.

From the joint PDF $p_{\mathrm{E}\Theta}(E, \theta)$, we can also obtain $p_\Theta(\theta)$, the PDF of the phase Θ, by integrating the joint PDF with respect to E,

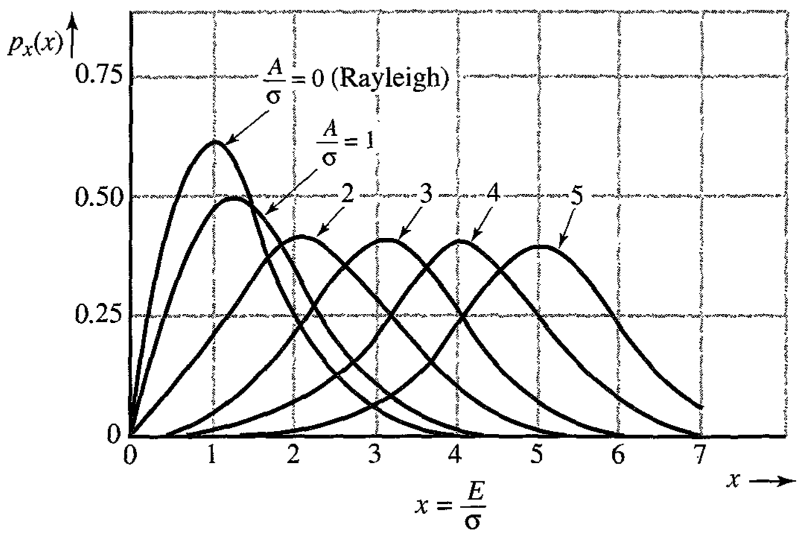


Figure 11.22 Rician PDF.

$$p_\Theta(\theta) = \int_0^\infty p_{E\Theta}(E, \theta)\, dE$$

Although the integration is straightforward, there are a number of involved steps, and for this reason it will not be repeated here. The final result is

$$p_\Theta(\theta) = \frac{1}{2\pi} e^{-A^2/2\sigma^2} \left\{ 1 + \frac{A}{\sigma}\sqrt{2\pi} \cos\theta e^{A^2 \cos^2\theta/2\sigma^2} \left[1 - Q\left(\frac{A\cos\theta}{\sigma} \right) \right] \right\} \tag{11.64d}$$

11.6 OPTIMUM FILTERING: WIENER-HOPF FILTER

When a desired signal is mixed with noise, the SNR can be improved by passing it through a filter that suppresses frequency components where the signal is weak but the noise is strong. The SNR improvement in this case can be explained qualitatively by considering a case of white noise mixed with a signal $\mathrm{m}(t)$ whose PSD decreases at high frequencies. If the filter attenuates higher frequencies more, the signal will be reduced—in fact, distorted. The distortion component $\mathrm{m}_\epsilon(t)$ may be considered an added noise. Thus, attenuation of higher frequencies will cause additional noise (from signal distortion), but, in compensation, it will reduce the channel noise, which is strong at high frequencies. Because at higher frequencies the signal has a small power content, the distortion component will be small compared to the reduction in channel noise, and the total noise may be smaller than before.

Let $H_{op}(\omega)$ be the optimum filter (Fig. 11.23a). This filter, not being ideal, will cause signal distortion. The distortion signal $\mathrm{m}_\epsilon(t)$ can be found from Fig. 11.23b. The distortion signal power N_D appearing at the output is given by

$$N_D = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega) |H_{op}(\omega) - 1|^2\, d\omega$$

where $S_{\mathrm{m}}(\omega)$ is the signal PSD at the input of the receiving filter. The channel noise power N_{ch} appearing at the filter output is given by

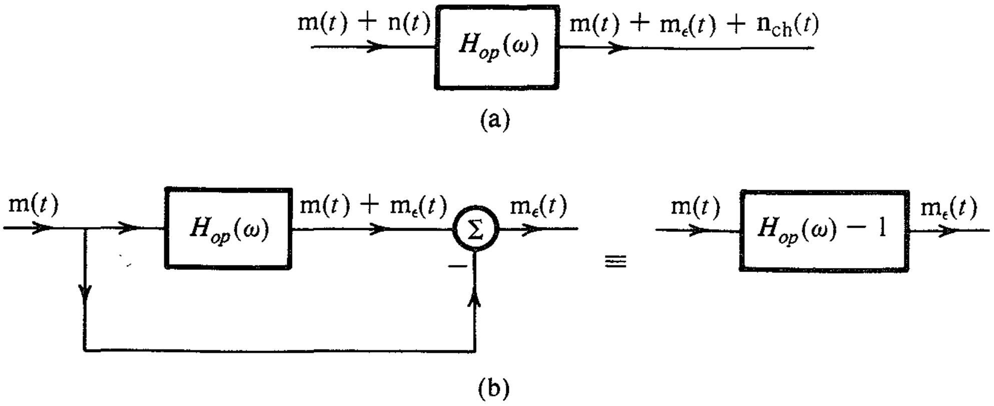


Figure 11.23 Wiener-Hopf filter calculations.

$$N_{ch} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{\rm n}(\omega)|H_{op}(\omega)|^2\,d\omega$$

where $S_{\rm n}(\omega)$ is the noise PSD appearing at the input of the receiving filter. The distortion component acts as a noise. Because the signal and the channel noise are incoherent, the total noise N_o at the receiving filter output is the sum of the channel noise N_{ch} and the distortion noise N_D,

$$\begin{aligned} N_o &= N_{ch} + N_D \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\left|H_{op}(\omega)\right|^2 S_{\rm n}(\omega) + \left|H_{op}(\omega) - 1\right|^2 S_{\rm m}(\omega)\right] d\omega \end{aligned} \tag{11.65a}$$

Using the fact that $|A+B|^2 = (A+B)(A^* + B^*)$, and noting that both $S_{\rm m}(\omega)$ and $S_{\rm n}(\omega)$ are real, Eq. (11.65a) can be rearranged as

$$N_o = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\left|H_{op}(\omega) - \frac{S_{\rm m}(\omega)}{S_{\rm r}(\omega)}\right|^2 S_{\rm r}(\omega) + \frac{S_{\rm m}(\omega)S_{\rm n}(\omega)}{S_{\rm r}(\omega)}\right] d\omega \tag{11.65b}$$

where $S_{\rm r}(\omega) = S_{\rm m}(\omega) + S_{\rm n}(\omega)$. The integrand on the right-hand side of Eq. (11.65b) is nonnegative. Moreover, it is a sum of two nonnegative terms. Hence, to minimize N_o, we must minimize each term. Because the second term $S_{\rm m}(\omega)S_{\rm n}(\omega)/S_{\rm r}(\omega)$ is independent of $H_{op}(\omega)$, only the first term can be minimized. From Eq. (11.65b) it is obvious that this term is minimum when

$$\begin{aligned} H_{op}(\omega) &= \frac{S_{\rm m}(\omega)}{S_{\rm r}(\omega)} \\ &= \frac{S_{\rm m}(\omega)}{S_{\rm m}(\omega) + S_{\rm n}(\omega)} \end{aligned} \tag{11.66a}$$

For this optimum choice, the output noise power N_o is given by

$$N_o = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{S_{\rm m}(\omega)S_{\rm n}(\omega)}{S_{\rm m}(\omega) + S_{\rm n}(\omega)}\,d\omega \tag{11.66b}$$

The optimum filter is known as the **Wiener-Hopf filter** in the literature. Equation (11.66a) shows that $H_{op}(\omega) \approx 1$ (no attenuation) when $S_m(\omega) \gg S_n(\omega)$. But when $S_m(\omega) \ll S_n(\omega)$, the filter has high attenuation. In other words, the optimum filter attenuates heavily the band where noise is relatively stronger. This causes some signal distortion, but at the same time it attenuates the noise more heavily so that the overall SNR is improved.

Comments on the Optimum Filter

If the SNR at the filter input is reasonably large—e.g., $S_m(\omega) > 100 S_n(\omega)$ (SNR of 20 dB)—the optimum filter [Eq. (11.66a)] in this case is practically an ideal filter, and N_o [Eq. (11.66b)] is given by

$$N_o \simeq \frac{1}{2\pi} \int_{-\infty}^{\infty} S_n(\omega)\, d\omega$$

Hence for a large input SNR, optimization yields insignificant improvement. The Wiener-Hopf filter is therefore practical only when the input SNR is small (large-noise case).

Another issue is the realizability of the optimum filter in Eq. (11.66a). Because $S_m(\omega)$ and $S_n(\omega)$ are both even functions of ω, the optimum filter $H_{op}(\omega)$ is an even function of ω. Hence, the unit impulse response $h_{op}(t)$ is an even function of t (see Prob. 3.1-1). This makes $h_{op}(t)$ noncausal and the filter unrealizable. As noted earlier, such a filter can be realized approximately if we are willing to tolerate some delay in the output. If delay cannot be tolerated, the derivation of $H_{op}(\omega)$ must be repeated with a realizability constraint. Note that the realizable optimum filter can never be superior to the unrealizable optimum filter [Eq. (11.66a)]. Thus, the filter in Eq. (11.66a) gives the upper bound on performance (output SNR). Discussion of realizable optimum filters can be readily found in the literature.[1,3]

EXAMPLE 11.13 A random process m(t) (the signal) is mixed with a white channel noise n(t). Given

$$S_m(\omega) = \frac{2\alpha}{\alpha^2 + \omega^2} \quad \text{and} \quad S_n(\omega) = \frac{\mathcal{N}}{2}$$

find the Wiener-Hopf filter to maximize the SNR. Find the resulting output noise power N_o.

From Eq. (11.66a),

$$\begin{aligned} H_{op}(\omega) &= \frac{4\alpha}{4\alpha + \mathcal{N}(\alpha^2 + \omega^2)} \\ &= \frac{4\alpha}{\mathcal{N}(\beta^2 + \omega^2)} \qquad \beta^2 = \frac{4\alpha}{\mathcal{N}} + \alpha^2 \end{aligned} \tag{11.67a}$$

Hence,

$$h_{op}(t) = \frac{2\alpha}{\mathcal{N}\beta} e^{-\beta|t|} \tag{11.67b}$$

Figure 11.24a shows $h_{op}(t)$. It is evident that this is an unrealizable filter. However, a delayed version (Fig. 11.24b) of this filter, that is, $h_{op}(t - t_0)$, is closely realizable if we make $t_0 \geq 3/\beta$ and eliminate the tail for $t < 0$ (Fig. 11.24c).

The output noise power N_o is [Eq. (11.66b)]

$$N_o = \frac{1}{2\pi} \int_0^{\infty} \frac{2\alpha}{\beta^2 + \omega^2}\, d\omega = \frac{\alpha}{\beta} = \frac{\alpha}{\sqrt{\alpha^2 + (4\alpha/\mathcal{N})}} \tag{11.68}$$

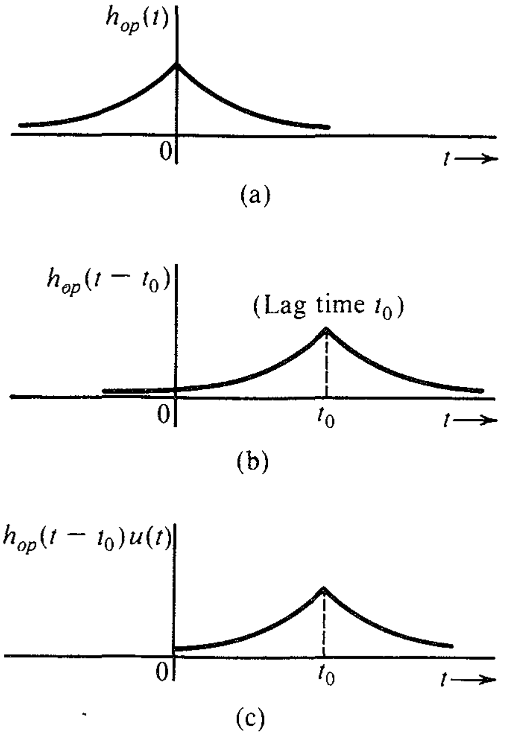


Figure 11.24 Close realization of an unrealizable filter using delay.

REFERENCES

1. B. P. Lathi, *An Introduction to Random Signals and Communication Theory*, International Textbook Co., Scranton, PA, 1968.
2. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
3. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed., McGraw-Hill, New York, 1984.
4. S. O. Rice, "Mathematical Analysis of Random Noise," *Bell. Syst. Tech. J.*, vol. 23, pp. 282–332, July 1944; vol. 24, pp. 46–156, Jan. 1945.

PROBLEMS

11.1-1 **(a)** Sketch the ensemble of the random process

$$x(t) = a\cos(\omega t + \Theta)$$

where ω and Θ are constants and a is an RV uniformly distributed in the range $(-A,\ A)$.

(b) Just by observing the ensemble, determine whether this is a stationary or a nonstationary process. Give your reasons.

11.1-2 Repeat Prob. 11.1-1 part **(a)** if a and Θ are constants but ω is an RV uniformly distributed in the range (0, 100).

11.1-3 **(a)** Sketch the ensemble of the random process

$$x(t) = at + b$$

where b is a constant and a is an RV uniformly distributed in the range $(-2,\ 2)$.

(b) Just by observing the ensemble, state whether this is a stationary or a nonstationary process.

11.1-4 Determine $\overline{x(t)}$ and $R_x(t_1,\ t_2)$ for the random process in Prob. 11.1-1, and determine whether this is a wide-sense stationary process.

11.1-5 Repeat Prob. 11.1-4 for the process $x(t)$ in Prob. 11.1-2.

11.1-6 Repeat Prob. 11.1-4 for the process $x(t)$ in Prob. 11.1-3.

11.1-7 Given a random process $x(t) = k$, where k is an RV uniformly distributed in the range $(-1,\ 1)$.

(a) Sketch the ensemble of this process.

(b) Determine $\overline{x(t)}$.

(c) Determine $R_x(t_1,\ t_2)$.

(d) Is the process wide-sense stationary?

(e) Is the process ergodic?

(f) If the process is wide-sense stationary, what is its power P_x [that is, its mean square value $\overline{x^2(t)}$]?

11.1-8 Repeat Prob. 11.1-7 for the random process

$$x(t) = a\cos(\omega_c t + \Theta)$$

where ω_c is a constant and a and Θ are independent RVs uniformly distributed in the ranges $(-1,\ 1)$ and $(0,\ 2\pi)$, respectively.

11.2-1 For each of the following functions, state whether it can be a valid PSD of a real random process.

(a) $\dfrac{\omega^2}{\omega^2 + 16}$ **(b)** $\dfrac{1}{\omega^2 - 16}$

(c) $\dfrac{\omega}{\omega^2 + 16}$ **(d)** $\delta(\omega) + \dfrac{1}{\omega^2 + 16}$

(e) $\delta(\omega + \omega_0) - \delta(\omega - \omega_0)$ **(f)** $j[\delta(\omega + \omega_0) + \delta(\omega - \omega_0)]$

(g) $\dfrac{j\omega^2}{\omega^2 + 16}$

11.2-2 Show that for a wide-sense stationary process $x(t)$,

(a) $R_x(0) \geq |R_x(\tau)| \qquad \tau \neq 0$

Hint: $\overline{(x_1 \pm x_2)^2} = \overline{x_1^2} + \overline{x_2^2} \pm \overline{2x_1x_2} \geq 0$. Let $x_1 = x(t_1)$ and $x_2 = x(t_2)$.

(b) $\lim_{\tau \to \infty} R_x(\tau) = \bar{x}^2$ *Hint:* As $\tau \to \infty$, x_1 and x_2 tend to become independent.

11.2-3 Show that if the PSD of a random process $x(t)$ is band-limited to B Hz, and if

$$R_x\left(\frac{n}{2B}\right) = \begin{cases} 1 & n = 0 \\ 0 & n = \pm 1,\ \pm 2,\ \pm 3,\ \ldots \end{cases}$$

then x(t) is a white bandlimited process; that is, $S_x(\omega) = k\,\text{rect}\,(\omega/4\pi B)$. *Hint:* Using the interpolation formula, reconstruct $R_x(\tau)$.

11.2-4 For the random binary process in Example 11.5 (Fig. 11.9a), determine $R_x(\tau)$ and $S_x(\omega)$ if the probability of transition (from 1 to -1 or vice versa) at each node is 0.6 instead of 0.5.

11.2-5 A wide-sense stationary white process m(t) band-limited to B Hz is sampled at the Nyquist rate. Each sample is transmitted by a basic pulse $p(t)$ multiplied by the sample value. This is a PAM signal. Show that the PSD of the PAM signal is $2BR_m(0)|P(\omega)|^2$. *Hint:* Use Eq. (11.29). Show that Nyquist samples a_k and a_{k+n} $(n \geq 1)$ are uncorrelated.

11.2-6 A duobinary line code proposed by Lender is a ternary scheme similar to bipolar, but requires only half the bandwidth of the latter. In this code, **0** is transmitted by no pulse, and **1** is transmitted by pulse $p(t)$ or $-p(t)$ using the following rule: A **1** is encoded by the same pulse as that used to encode the preceding **1** if the two **1**'s are separated by an even number of **0**'s. It is encoded by the negative of the pulse used to encode the preceding **1** if the two **1**'s are separated by an odd number of **0**'s. Random binary digits are transmitted every T_b seconds. Assuming $P(0) = P(1) = 0.5$, show that

$$S_y(\omega) = \frac{|P(\omega)|^2}{T_b}\cos^2\left(\frac{\omega T_b}{2}\right)$$

Find $S_y(\omega)$ if $p(t)$, the basic pulse used, is a half-width rectangular pulse rect $(2t/T_b)$.

11.2-7 Determine $S_y(\omega)$ for polar signaling if $P(1) = Q$ and $P(0) = 1 - Q$.

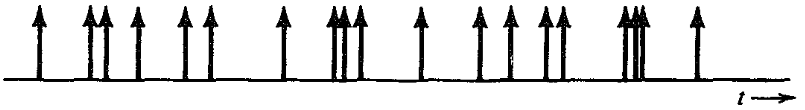


Figure P11.2-8

11.2-8 An impulse noise x(t) can be modeled by a sequence of unit impulses located at random instants (Fig. P11.2-8). There are an average of α impulses per second, and the location of any impulse is independent of the locations of other impulses. Show that $R_x(\tau) = \alpha\,\delta(\tau) + \alpha^2$.

11.2-9 Repeat Prob. 11.2-8 if the impulses are equally likely to be positive and negative.

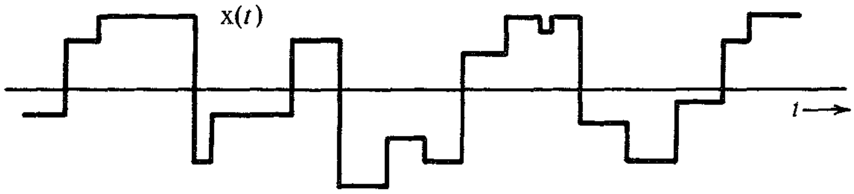


Figure P11.2-10

11.2-10 A sample function of a random process x(t) is shown in Fig. P11.2-10. The signal x(t) changes abruptly in amplitude at random instants. There are an average of β amplitude changes (or shifts) per second. The probability that there will be no amplitude shift in τ seconds is given by $P_0(\tau) = e^{-\beta\tau}$. The amplitude after a shift is independent of the amplitude before the shift. The amplitudes are randomly distributed, with a PDF $p_{\mathrm{x}}(x)$. Show that

$$R_{\mathrm{x}}(\tau) = \overline{\mathrm{x}^2} e^{-\beta|\tau|} \qquad \text{and} \qquad S_{\mathrm{x}}(\omega) = \frac{2\beta\overline{\mathrm{x}^2}}{\beta^2 + \omega^2}$$

This process represents a model for thermal noise.[1]

11.3-1 Show that for jointly wide-sense stationary, real, random processes $\mathrm{x}(t)$ and $\mathrm{y}(t)$,

$$|R_{\mathrm{xy}}(\tau)| \leq [R_{\mathrm{x}}(0) R_{\mathrm{y}}(0)]^{1/2}$$

Hint: For any real number a, $\overline{(a\mathrm{x} - \mathrm{y})^2} \geq 0$.

11.3-2 If $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are two incoherent random processes, and two new processes $\mathrm{u}(t)$ and $\mathrm{v}(t)$ are formed as follows:

$$\mathrm{u}(t) = \mathrm{x}(t) + \mathrm{y}(t) \qquad \mathrm{v}(t) = 2\mathrm{x}(t) + 3\mathrm{y}(t)$$

find $R_{\mathrm{u}}(\tau)$, $R_{\mathrm{v}}(\tau)$, $R_{\mathrm{uv}}(\tau)$, and $R_{\mathrm{vu}}(\tau)$ in terms of $R_{\mathrm{x}}(\tau)$ and $R_{\mathrm{y}}(\tau)$.

11.3-3 Two random processes $\mathrm{x}(t)$ and $\mathrm{y}(t)$ are

$$\mathrm{x}(t) = A\cos(\omega_0 t + \varphi) \qquad \text{and} \qquad \mathrm{y}(t) = B\cos(n\omega_0 t + n\varphi)$$

where n = integer $\neq 1$ and A, B, and ω_0 are constants and φ is an RV uniformly distributed in the range $(0, 2\pi)$. Show that the two processes are incoherent.

11.3-4 A sample function of a periodic random process $\mathrm{x}(t)$ is shown in Fig. P11.3-4. The interval b where the first pulse begins is an RV uniformly distributed in the range $(0, T_b)$. If a sample function with $\mathrm{b} = 0$ is expressed by a compact trigonometric Fourier series,

$$\mathrm{x}(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

then show that

$$R_{\mathrm{x}}(\tau) = C_0^2 + \frac{1}{2}\sum_{n=1}^{\infty} C_n^2 \cos n\omega_0\tau \qquad \omega_0 = \frac{2\pi}{T_0}$$

Find the PSD $S_{\mathrm{x}}(\omega)$. *Hint*:

$$\mathrm{x}(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos[n\omega_0(t - \mathrm{b}) + \theta_n]$$

where b is an RV uniformly distributed in the range $(0, T_0)$. Now use the results in Prob. 11.3-3.

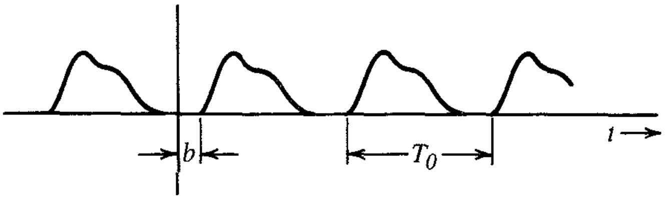


Figure P11.3-4

11.4-1 A simple RC circuit has two resistors R_1 and R_2 in parallel (Fig. P11.4-1a). Calculate the rms value of the thermal noise voltage v_o across the capacitor in two ways:

(a) Consider resistors R_1 and R_2 as two separate resistors, with respective thermal noise voltages of PSD $2kTR_1$ and $2kTR_2$ (Fig. P11.4-1b). Note that the two sources are independent.

(b) Consider the parallel combination of R_1 and R_2 as a single resistor of value $R_1R_2/(R_1 + R_2)$, with its thermal-noise voltage source of PSD $2kTR_1R_2/(R_1 + R_2)$ (Fig. P11.4- 1c). Comment.

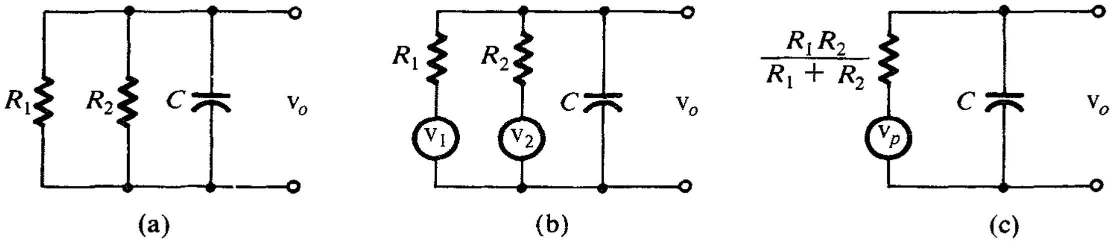


Figure P11.4-1

11.4-2 Show that $R_{xy}(\tau)$, the crosscorrelation function of the input process x(t) and the output process y(t) in Fig. 11.12, is

$$R_{xy}(\tau) = h(\tau) * R_x(\tau) \qquad \text{and} \qquad S_{xy}(\omega) = H(\omega)S_x(\omega)$$

Hence, show that for the thermal noise n(t) and the output $v_o(t)$ in Fig. 11.13 (Example 11.9),

$$S_{nv_o}(\omega) = \frac{2kTR}{1 + j\omega RC} \qquad \text{and} \qquad R_{nv_o}(\tau) = \frac{2kT}{C}e^{-\tau/RC}\,u(\tau)$$

11.4-3 A shot noise is similar to impulse noise described in Prob. 11.2-8 except that instead of random impulses, we have pulses of finite width. If we replace each impulse in Fig. P11.2-8 by a pulse $h(t)$ whose width is large compared to $1/\alpha$, so that there is a considerable overlapping of pulses, we get shot noise. The result of pulse overlapping is that the signal looks like a continuous random signal, as shown in Fig. P11.4-3.

(a) Derive the autocorrelation function and the PSD of such a random process. *Hint:* Shot noise results from passing impulse noise through a suitable filter. First derive the PSD of the shot noise and then obtain the autocorrelation function from the PSD. The answers will be in terms of α, $h(t)$, or $H(\omega)$.

(b) The shot noise in transistors can be modeled by

$$h(t) = \frac{q}{T}e^{-t/T}u(t)$$

where q is the charge on an electron and T is the electron transit time. Determine and sketch the autocorrelation function and the PSD of the transistor shot noise.

Figure P11.4-3

11.5-1 A white process of PSD $\mathcal{N}/2$ is transmitted through a bandpass filter $H(\omega)$ (Fig. P11.5-1). Represent the filter output $n(t)$ in terms of quadrature components, and determine $S_{n_c}(\omega)$, $S_{n_s}(\omega)$, $\overline{n_c^2}$, $\overline{n_s^2}$, and $\overline{n^2}$ when the center frequency used in this representation is 100 kHz (that is, $\omega_c = 200\pi \times 10^3$).

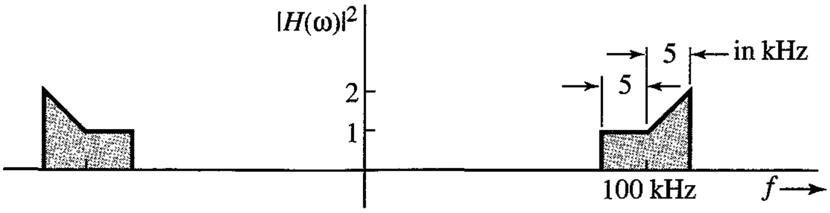


Figure P11.5-1

11.5-2 Repeat Prob. 11.5-1 if the center frequency ω_c used in the representation is not a true center frequency. Consider three cases: **(a)** $f_c = 105$ kHz; **(b)** $f_c = 95$ kHz; **(c)** $f_c = 90$ kHz.

11.5-3 A random process $x(t)$ with the PSD shown in Fig. P11.5-3a is passed through a bandpass filter (Fig. 11.5-3b). Determine the PSDs and mean square values of the quadrature components of the output process. Assume the center frequency in the representation to be 0.5 MHz.

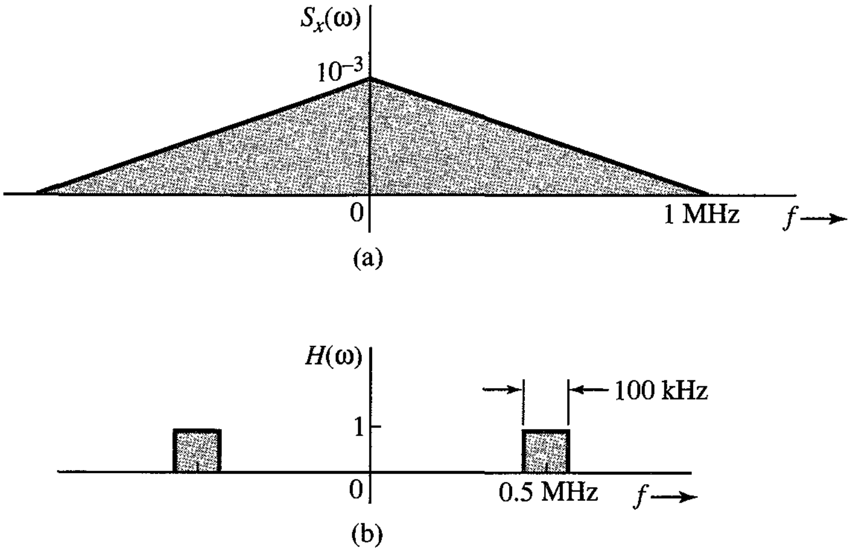


Figure P11.5-3

11.5-4 A signal process $m(t)$ is mixed with a channel noise $n(t)$. The respective PSDs are

$$S_m(\omega) = \frac{6}{9+\omega^2} \quad \text{and} \quad S_n(\omega) = 6$$

(a) Find the optimum Wiener-Hopf filter.

(b) Sketch its unit impulse response.

(c) Estimate the amount of delay necessary to make this filter closely realizable.

(d) Compute the noise power at the input and the output of the filter.

(e) What is the SNR improvement realized by the use of this filter? *Hint*: Note that because of the special way in which the problem of optimum filtering is formulated, the signal powers at the input and the output of the filter are identical (see Fig. 11.23).

11.5-5 Repeat Prob. 11.5-4 if

$$S_{\mathrm{m}}(\omega) = \frac{4}{4+\omega^2} \qquad \text{and} \qquad S_{\mathrm{n}}(\omega) = \frac{32}{64+\omega^2}$$

12 BEHAVIOR OF ANALOG SYSTEMS IN THE PRESENCE OF NOISE

In this chapter, we analyze the behavior of analog communication systems in the presence of noise to facilitate the comparison of various systems.

Figure 12.1 shows a schematic of a communication system. A certain signal power S_T is transmitted over a channel.* The transmitted signal is corrupted by channel noise during transmission. We shall assume channel noise to be additive. The channel will attenuate (and may also distort) the signal. At the receiver input, we have a signal mixed with noise. The signal and noise powers at the receiver input are S_i and N_i, respectively. The receiver processes (filters, demodulates, etc.) the signal to yield the desired signal plus noise. The signal and noise powers at the receiver output are S_o and N_o, respectively. In analog systems, the quality of the received signal is determined by S_o/N_o, the output SNR. Hence, we shall focus our attention on this parameter. But S_o/N_o can be increased as much as desired simply by increasing the transmitted power S_T. In practice, however, the maximum value of S_T is limited by other considerations, such as transmitter cost, channel capability, interference with other channels, and so on. Hence, the value of S_o/N_o for a given transmitted power is an appropriate figure of merit in an analog communication system. In practice, it is more convenient to deal with the received power S_i rather than the transmitted power S_T. From Fig. 12.1, it is apparent that S_i is proportional to S_T. Hence, the value of S_o/N_o for a given S_i will serve equally well as a figure of merit.

12.1 BASEBAND SYSTEMS

In baseband systems, the signal is transmitted directly without any modulation. This mode of communication is suitable over a pair of wires, optical fiber, or coaxial cables. It is mainly used in short-haul links. The study of baseband systems is important because many of the basic concepts and parameters encountered in baseband systems are carried over directly to modulated systems. Second, baseband systems serve as a basis against which other systems may be compared.

---

* Here the channel is used in the sense of a transmission medium.

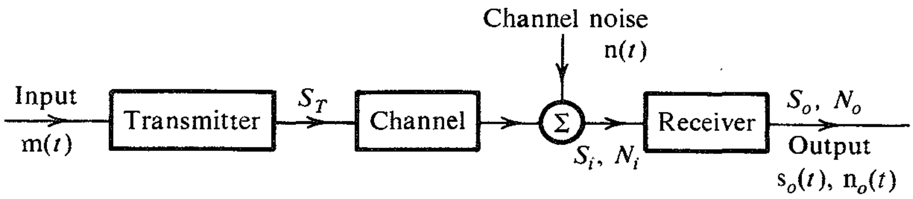

Channel noise
n(t)
Input
m(t)
Transmitter
S_T
Channel
Σ
S_i, N_i
Receiver
S_o, N_o
Output
$s_o(t)$, $n_o(t)$

Figure 12.1 Communication system model.

For a baseband system, the transmitter and the receiver are ideal baseband filters (Fig. 12.2). The low-pass filter $H_p(\omega)$ at the transmitter limits the input signal spectrum to a given bandwidth. The low-pass filter $H_d(\omega)$ at the receiver eliminates the out-of-band noise and other channel interference. These filters can also serve an additional purpose, that of preemphasis and deemphasis, which optimizes the signal-to-noise ratio (SNR) at the receiver (or minimizes the channel noise interference).

The baseband signal m(t) is assumed to be a zero mean, wide-sense stationary random process band-limited to B Hz. To begin with, we shall consider the case of ideal low-pass (or baseband) filters with bandwidth B at the transmitter and the receiver (Fig. 12.2). The channel is assumed to be distortionless. For this case,

$$S_o = S_i \tag{12.1a}$$

and

$$N_o = 2\int_0^B S_n(\omega)\,df \tag{12.1b}$$

where $S_n(\omega)$ is the PSD of the channel noise. For the case of a white noise, $S_n(\omega) = \mathcal{N}/2$, and

$$N_o = 2\int_0^B \frac{\mathcal{N}}{2}\,df = \mathcal{N}B \tag{12.1c}$$

and

$$\frac{S_o}{N_o} = \frac{S_i}{\mathcal{N}B} \tag{12.1d}$$

We define a parameter γ as

$$\gamma = \frac{S_i}{\mathcal{N}B} \tag{12.2}$$

From Eqs. (12.1d) and (12.2) we have

$$\frac{S_o}{N_o} = \gamma \tag{12.3}$$

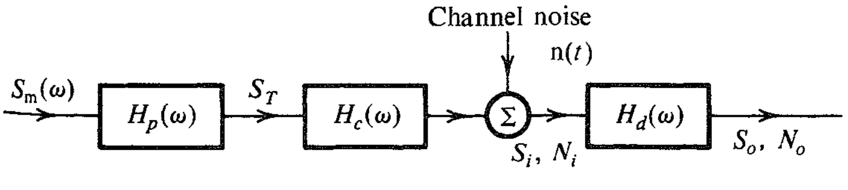

Channel noise
n(t)
$S_m(\omega)$
$H_p(\omega)$
S_T
$H_c(\omega)$
Σ
S_i, N_i
$H_d(\omega)$
S_o, N_o

Figure 12.2 Baseband system.

The parameter γ is directly proportional to S_i and, therefore, directly proportional to S_T. Hence, a given S_T (or S_i) implies a given γ. Equation (12.3) is precisely the result we are looking for. It gives the receiver output SNR for a given S_T (or S_i).

The value of the SNR in Eq. (12.3) will serve as a standard against which the output SNR of other systems will be measured.

The power, or the mean square value, of $m(t)$ is $\overline{m^2}$, given by

$$\overline{m^2} = 2\int_0^B S_m(\omega)\,df \tag{12.4}$$

In analog signals, the SNR is basic in specifying the signal quality. For voice signals, an SNR of 5 to 10 dB at the receiver implies a barely intelligible signal. Telephone quality signals have an SNR of 25 to 35 dB, whereas for television, an SNR of 45 to 55 dB is required.

12.2 AMPLITUDE-MODULATED SYSTEMS

We shall analyze DSB-SC, SSB-SC, and AM systems separately.

DSB-SC

A basic DSB-SC system is shown in Fig. 12.3.* The modulated signal is a bandpass signal centered at ω_c with a bandwidth $2B$. The channel noise is assumed to be additive. The channel and the filters in Fig. 12.3 are assumed to be ideal.

Let S_i and S_o represent the useful signal powers at the input and the output of the demodulator, and let N_o represent the noise power at the demodulator output. The signal at the demodulator input is $\sqrt{2}\,m(t)\cos\omega_c t + n(t)$, where $n(t)$ is the bandpass channel noise. Its spectrum is centered at ω_c and has a bandwidth $2B$. The input signal power S_i is the power of the modulated signal† $\sqrt{2}\,m(t)\cos\omega_c t$. From Eq. (11.20c),

$$S_i = \overline{[\sqrt{2}\,m(t)\cos\omega_c t]^2} = (\sqrt{2})^2\overline{[m(t)\cos\omega_c t]^2} = \overline{m^2(t)} = \overline{m^2} \tag{12.5}$$

The reader may now appreciate our use of $\sqrt{2}\cos\omega_c t$ (rather than $\cos\omega_c t$) in the modulator (Fig. 12.3). This was done to make the received power equal to that in the baseband system in order to facilitate comparison. We shall use a similar artifice in our analysis of the SSB system.

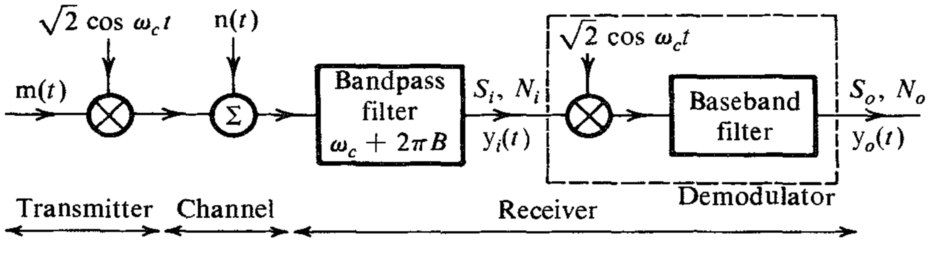


Figure 12.3 DSB-SC system.

---

* The use of an input bandpass filter in the receiver may appear redundant because the out-of-band noise components will be suppressed by the final baseband filter. In practice, an input filter is useful because by removing the out-of-band noise, it reduces the probability of nonlinear distortion from overload effects.

† The modulated signal also has a random phase Θ, which is uniformly distributed in the range $(0, 2\pi)$. This random phase [which is independent of $m(t)$] does not affect the final results and, hence, is ignored in this discussion.

To determine the output powers S_o and N_o, we note that the signal at the demodulator input is

$$y_i(t) = \sqrt{2}\, m(t) \cos \omega_c t + n(t)$$

Because $n(t)$ is a bandpass signal centered at ω_c, we can express it in terms of quadrature components, as in Eq. (11.48). This gives

$$y_i(t) = \left[\sqrt{2}\, m(t) + n_c(t)\right] \cos \omega_c t + n_s(t) \sin \omega_c t$$

When this signal is multiplied by $\sqrt{2} \cos \omega_c t$ (synchronous demodulation) and then low-pass filtered, the terms $m(t)\ \cos\ 2\omega_c t$ and $m(t)\ \sin\ 2\omega_c t$ are suppressed. The resulting demodulator output $y_o(t)$ is

$$y_o(t) = m(t) + \frac{1}{\sqrt{2}} n_c(t)$$

Hence,

$$S_o = \overline{m^2} = S_i \tag{12.6a}$$

$$N_o = \frac{1}{2} \overline{n_c^2(t)} \tag{12.6b}$$

For white noise with power density $\mathcal{N}/2$, we have [Eq. (11.50b)]

$$\overline{n_c^2(t)} = \overline{n^2(t)} = 2\mathcal{N}B$$

and

$$N_o = \mathcal{N}B \tag{12.7}$$

Hence, from Eqs. (12.6a) and (12.7) we have

$$\frac{S_o}{N_o} = \frac{S_i}{\mathcal{N}B} = \gamma \tag{12.8}$$

Comparison of Eqs. (12.8) and (12.3) shows that for a fixed transmitted power (which also implies a fixed signal power at the demodulator input), the SNR at the demodulator output is the same for the baseband and the DSB-SC systems. Moreover, quadrature multiplexing in DSB-SC can render its bandwidth requirement identical to that of baseband systems. Thus, theoretically, baseband and DSB-SC systems have identical capabilities.

SSB-SC

An SSB-SC system is shown in Fig. 12.4. The SSB signal* $\varphi_{\text{SSB}}(t)$ can be expressed as [see Eq. (4.17c)]

$$\varphi_{\text{SSB}}(t) = m(t) \cos \omega_c t + m_h(t) \sin \omega_c t \tag{12.9}$$

The spectrum of $\varphi_{\text{SSB}}(t)$ is shown in Fig. 4.15. This signal can be obtained (Fig. 12.4) by multiplying $m(t)$ by $2\ \cos\ \omega_c t$ and then suppressing the unwanted sideband. The power of the modulated signal $2\, m(t)\ \cos\ \omega_c t$ is $2\,\overline{m^2}$ [four times the power of $m(t) \cos \omega_c t$]. Suppression of one sideband halves the power. Hence S_i, the power of $\varphi_{\text{SSB}}(t)$, is

$$S_i = \overline{m^2} \tag{12.10}$$

---

* This is LSB. The discussion is valid for USB as well.

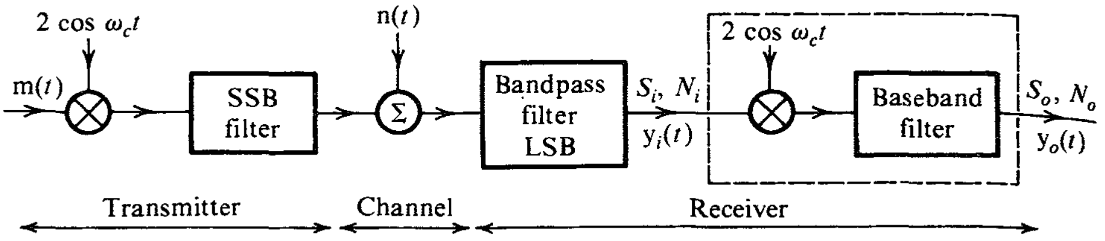


Figure 12.4 SSB-SC system.

Expressing the channel bandpass noise in terms of quadrature components as in Eq. (11.52), the signal at the detector input, $y_i(t)$, is

$$y_i(t) = [m(t) + n_c(t)] \cos \omega_c t + [m_h(t) + n_s(t)] \sin \omega_c t$$

This signal is multiplied by $2 \cos \omega_c t$ (synchronous demodulation) and then low-pass filtered to yield the demodulator output

$$y_o(t) = m(t) + n_c(t)$$

Hence,

$$S_o = \overline{m^2} = S_i$$

$$N_o = \overline{n_c^2} \tag{12.11}$$

We have already found $\overline{n_c^2}$ for the SSB channel noise (lower sideband) in Eq. (11.54b) as

$$N_o = \overline{n_c^2} = \mathcal{N}B$$

Thus,

$$\frac{S_o}{N_o} = \frac{S_i}{\mathcal{N}B} = \gamma \tag{12.12}$$

This shows that baseband, DSB-SC, and SSB-SC systems perform identically in terms of resource utilization. All of them yield the same output SNR for a given transmitted power and transmission bandwidth.

EXAMPLE 12.1 In a DSB-SC system, the carrier frequency is $f_c = 500$ kHz, and the modulating signal m(t) has a uniform PSD band-limited to 4 kHz. The modulated signal is transmitted over a distortionless channel with a noise PSD $S_n(\omega) = 1/(\omega^2 + a^2)$, where $a = 10^6\pi$. The useful signal power at the receiver input is 1 μW. The received signal is bandpass filtered, multiplied by $2 \cos \omega_c t$, and then low-pass filtered to obtain the output $s_o(t) + n_o(t)$. Determine the output SNR.

If the received signal is km(t) cos $\omega_c t$, the demodulator input is $[km(t) + n_c(t)] \cos \omega_c t + n_s(t) \sin \omega_c t$. When this is multiplied by $2 \cos \omega_c t$ and low-pass filtered, the output is

$$s_o(t) + n_o(t) = km(t) + n_c(t) \tag{12.13}$$

Hence,

$$S_o = k^2\overline{m^2} \qquad \text{and} \qquad N_o = \overline{n_c^2}$$

But the power of the received signal $k\mathrm{m}(t)\cos\omega_c t$ is 1 μW. Hence,

$$\frac{k^2\overline{\mathrm{m}^2}}{2} = 10^{-6}$$

and

$$S_o = k^2\overline{\mathrm{m}^2} = 2 \times 10^{-6}$$

To compute $\overline{\mathrm{n}_c^2}$, we use Eq. (11.45c):

$$\overline{\mathrm{n}_c^2} = \overline{\mathrm{n}^2}$$

where $\overline{\mathrm{n}^2}$ is the power of the incoming bandpass noise of bandwidth 8 kHz centered at 500 kHz; that is,

$$\begin{aligned}\overline{\mathrm{n}^2} &= 2\int_{496,000}^{504,000} \frac{1}{\omega^2 + a^2}\,df \qquad a = 10^6\pi \\ &= \frac{1}{\pi}\int_{(2\pi)496,000}^{(2\pi)504,000} \frac{1}{\omega^2 + a^2}\,d\omega = \frac{1}{\pi a}\tan^{-1}\frac{\omega}{a}\bigg|_{(2\pi)496,000}^{(2\pi)504,000} \\ &= 8.25 \times 10^{-10} = N_o\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{S_o}{N_o} &= \frac{2 \times 10^{-6}}{8.25 \times 10^{-10}} = 2.42 \times 10^3 \\ &= 33.83\,\mathrm{dB}\end{aligned}$$

AM

AM signals can be demodulated synchronously or by envelope detection. The former is of theoretical interest only. It is useful, however, for comparing the noise performance of the envelope detector. For this reason, we shall consider both of these methods.

Coherent, or Synchronous, Demodulation of AM: Coherent AM is identical to DSB-SC in every respect except for the additional carrier. If the received signal $\sqrt{2}[A+\mathrm{m}(t)]\cos\omega_c t$ is multiplied by $\sqrt{2}\cos\omega_c t$, the demodulator output is $\mathrm{m}(t)$. Hence,

$$S_o = \overline{\mathrm{m}^2}$$

The output noise will be exactly the same as that in DSB-SC [Eq. (12.7)]:

$$N_o = \overline{\mathrm{n}_o^2} = \mathcal{N}B$$

The received signal is $\sqrt{2}[A + \mathrm{m}(t)]\cos\omega_c t$. Hence,

$$\begin{aligned}S_i &= (\sqrt{2})^2\frac{\overline{[A + \mathrm{m}(t)]^2}}{2} \\ &= \overline{[A + \mathrm{m}(t)]^2} \\ &= A^2 + \overline{\mathrm{m}^2(t)} + 2A\,\overline{\mathrm{m}(t)}\end{aligned}$$

Because $\mathrm{m}(t)$ is assumed to have zero mean,

$$S_i = A^2 + \overline{\mathrm{m}^2(t)}$$

and

$$\begin{aligned}\frac{S_o}{N_o} &= \frac{\overline{\mathrm{m}^2}}{\mathcal{N}B} \\ &= \frac{\overline{\mathrm{m}^2}}{A^2+\overline{\mathrm{m}^2}}\frac{S_i}{\mathcal{N}B} \\ &= \frac{\overline{\mathrm{m}^2}}{A^2+\overline{\mathrm{m}^2}}\gamma \end{aligned} \tag{12.14}$$

If $\mathrm{m}(t)_{\max} = m_p$, then $A \geq m_p$. For the maximum SNR, $A = m_p$, and

$$\begin{aligned}\left(\frac{S_o}{N_o}\right)_{\max} &= \frac{\overline{\mathrm{m}^2}}{m_p^2+\overline{\mathrm{m}^2}}\gamma \\ &= \frac{1}{(m_p^2/\overline{\mathrm{m}^2}+1)}\gamma \end{aligned} \tag{12.15a}$$

Because $(m_p^2/\overline{\mathrm{m}^2}) \geq 1$,

$$\frac{S_o}{N_o} \leq \frac{\gamma}{2} \tag{12.15b}$$

It can be seen that the SNR in AM is at least 3 dB (and usually about 6 dB in practice) worse than that in DSB-SC and SSB-SC (depending on the modulation index and the signal waveform). For example, when $\mathrm{m}(t)$ is sinusoidal, $m_p^2/\overline{\mathrm{m}^2} = 2$, and AM requires three times as much power (4.77 dB) as that needed for DSB-SC or SSB-SC.

In many communication systems the transmitter is limited by peak power rather than average power transmitted. In such a case, AM fares even worse. It can be shown (see Prob. 12.2-6) that in tone modulation, for a fixed peak power transmitted, the output SNR of AM is 6 dB below that of DSB-SC and 9 dB below that of SSB-SC. These results are valid under conditions most favorable to AM, that is, with modulation index $\mu = 1$. For $\mu < 1$, AM would be worse than this. For this reason, volume compression and peak limiting are generally used in AM transmission in order to have full modulation most of the time.

AM Envelope Detection: Assuming the received signal to be $[A + \mathrm{m}(t)] \cos \omega_c t$, the demodulator input is

$$\mathrm{y}_i(t) = [A + \mathrm{m}(t)] \cos \omega_c t + \mathrm{n}(t)$$

Using the quadrature component representation for $\mathrm{n}(t)$, we have

$$\mathrm{y}_i(t) = [A + \mathrm{m}(t) + \mathrm{n}_c(t)] \cos \omega_c t + \mathrm{n}_s(t) \sin \omega_c t \tag{12.16a}$$

The desired signal at the demodulator input is $[A + \mathrm{m}(t)] \cos \omega_c t$. Hence, the signal power S_i

is [see Eq. (11.20c)]

$$S_i = \frac{\overline{[A + \mathrm{m}(t)]^2}}{2} = \frac{A^2 + \overline{\mathrm{m}^2}}{2}$$

To find the envelope of $\mathrm{y}_i(t)$, we rewrite Eq. (12.16a) in polar form as

$$\mathrm{y}_i(t) = \mathrm{E}_i(t) \cos[\omega_c t + \Theta_i(t)] \tag{12.16b}$$

where the envelope $\mathrm{E}_i(t)$ is [see Eq. (3.40)]

$$\mathrm{E}_i(t) = \sqrt{[A + \mathrm{m}(t) + \mathrm{n}_c(t)]^2 + \mathrm{n}_s^2(t)} \tag{12.16c}$$

The envelope detector output is $\mathrm{E}_i(t)$ [Eq. (12.16c)]. We shall consider two extreme cases: small noise and large noise.

1. Small-Noise Case: If $[A + \mathrm{m}(t)] \gg \mathrm{n}(t)$ for almost all t, then $[A + \mathrm{m}(t)] \gg \mathrm{n}_c(t)$ and $\mathrm{n}_s(t)$ for almost all t.* In this case $\mathrm{E}_i(t)$ in Eq. (12.16c) can be approximated by $[A + \mathrm{m}(t) + \mathrm{n}_c(t)]$,

$$\mathrm{E}_i(t) \simeq A + \mathrm{m}(t) + \mathrm{n}_c(t)$$

The dc component A of the envelope detector output E_i is blocked by a capacitor, yielding $\mathrm{m}(t)$ as the useful signal and $\mathrm{n}_c(t)$ as the noise. Hence,

$$S_o = \overline{\mathrm{m}^2}$$

and from Eq. (11.50b),

$$N_o = \overline{\mathrm{n}_c^2(t)} = 2\mathcal{N}B$$

and

$$\begin{aligned} \frac{S_o}{N_o} &= \frac{\overline{\mathrm{m}^2}}{2\mathcal{N}B} \\ &= \frac{\overline{\mathrm{m}^2}}{A^2 + \overline{\mathrm{m}^2}} \frac{S_i}{\mathcal{N}B} \\ &= \frac{\overline{\mathrm{m}^2}}{A^2 + \overline{\mathrm{m}^2}} \gamma \end{aligned} \tag{12.17}$$

which is identical to the result for AM with synchronous demodulation [Eq. (12.14)]. Therefore for AM, when the noise is small compared to the signal, the performance of the envelope detector is identical to that of the synchronous detector.

2. Large-Noise Case: In this case $\mathrm{n}(t) \gg [A + \mathrm{m}(t)]$. Hence, $\mathrm{n}_c(t)$ and $\mathrm{n}_s(t) \gg [A + \mathrm{m}(t)]$ for almost all t. Under this condition Eq. (12.16c) becomes

---

* Here we use the term "almost all t" because $\mathrm{n}_c(t)$ and $\mathrm{n}_s(t)$ are both gaussian (amplitude range $-\infty$ to ∞), and in some instances $\mathrm{n}_c(t)$ or $\mathrm{n}_s(t)$ or both will exceed $A + \mathrm{m}(t)$, no matter how large $A + \mathrm{m}(t)$ is. For large signals, however, this occurs only over relatively short time intervals.

$$
\begin{aligned}
E_i(t) &\simeq \sqrt{n_c^2(t) + n_s^2(t) + 2n_c(t)[A + m(t)]} \\
&= E_n(t)\sqrt{1 + \frac{2[A + m(t)]}{E_n(t)} \cos \Theta_n(t)}
\end{aligned}
$$

where $E_n(t)$ and $\Theta_n(t)$, the envelope and the phase of the noise $n(t)$, are [see Eqs. (3.40) and Fig. 12.5a]

$$E_n(t) = \sqrt{n_c^2(t) + n_s^2(t)} \tag{12.18a}$$

$$\Theta_n(t) = -\tan^{-1}\left[\frac{n_s(t)}{n_c(t)}\right] \tag{12.18b}$$

$$n_c(t) = E_n(t) \cos \Theta_n(t) \tag{12.18c}$$

$$n_s(t) = E_n(t) \sin \Theta_n(t) \tag{12.18d}$$

Because $E_n(t) \gg A + m(t)$, $E_i(t)$ may be further approximated as

$$
\begin{aligned}
E_i(t) &\simeq E_n(t)\left[1 + \frac{A + m(t)}{E_n(t)} \cos \Theta_n(t)\right] \\
&= E_n(t) + [A + m(t)] \cos \Theta_n(t)
\end{aligned}
\tag{12.19}
$$

A glance at Eq. (12.19) shows that the output contains no term proportional to $m(t)$. The signal $m(t) \cos \Theta_n(t)$ represents $m(t)$ multiplied by a time-varying function (actually a noise signal) $\cos \Theta_n(t)$ and, hence, is of no use in recovering $m(t)$. In all previous cases, the output signal contained a term of the form $am(t)$, where a was constant. Furthermore, the output noise was additive (even for envelope detection with small noise). In Eq. (12.19), the noise is multiplicative. In this situation the useful signal is badly mutilated. This is the threshold phenomenon, where the signal quality at the output undergoes disproportionately rapid deterioration when the input noise increases beyond a certain level (i.e., when γ drops below a certain value.)

Calculation of the SNR for the intermediate case (transition region) is quite complex.[1] Here we shall state the final results only:

$$\frac{S_o}{N_o} \simeq 0.916A^2\overline{m^2}\,\gamma^2$$

Figure 12.5b shows the plot of S_o/N_o as a function of γ for AM with synchronous detection and AM with envelope detection. The threshold effect is clearly seen from this figure. The threshold occurs when γ is on the order of 10 or less. For a reasonable-quality AM signal, γ should be on the order of 1000 (30 dB), and the threshold is rarely a limiting condition in practical cases.

EXAMPLE 12.2 Find γ_{thresh}, the value of γ at the threshold, in tone-modulated AM with $\mu = 1$ if the onset of the threshold is when $E_n > A$ with probability 0.01, E_n being the noise envelope.

Because $E_n = \sqrt{n_c^2 + n_s^2}$, where both n_c and n_s are gaussian with variance σ_n^2, according to Eq. (10.54a) E_n has a Rayleigh PDF with variance σ_n^2, and

$$
\begin{aligned}
P(\mathrm{E_n} \geq A) &= \int_A^\infty \frac{E_\mathrm{n}}{\sigma_\mathrm{n}^2} e^{-En^2/2\sigma_\mathrm{n}^2}\, dE_\mathrm{n} \\
&= e^{-A^2/2\sigma_\mathrm{n}^2} = 0.01
\end{aligned}
$$

Hence, at the onset of the threshold,

$$\frac{A^2}{2\sigma_\mathrm{n}^2} = 4.605$$

The variance σ_n^2 of the bandpass noise of PSD $\mathcal{N}/2$ and the bandwidth $2B$ is $2\mathcal{N}B$. Hence, at the onset of the threshold,

$$\frac{A^2}{4\mathcal{N}B} = 4.605$$

For tone modulation,

$$
\begin{aligned}
\mathrm{m}(t) &= \mu A \cos(\omega_\mathrm{m} t + \Theta) \\
&= A \cos(\omega_\mathrm{m} t + \Theta) \qquad \mu = 1
\end{aligned}
$$

and

$$S_i = \frac{A^2 + \overline{\mathrm{m}^2}}{2} = \frac{A^2 + 0.5A^2}{2} = \frac{3A^2}{4}$$

Hence,

$$\gamma_{\text{thresh}} = \frac{S_i}{\mathcal{N}B} = \frac{3A^2}{4\mathcal{N}B} = 3(4.605) = 13.8 \quad \text{(or 12.4 dB)}$$

12.3 ANGLE-MODULATED SYSTEMS

A block diagram of an angle-modulated system is shown in Fig. 12.6. The angle-modulated (or exponentially modulated) carrier $\varphi_{\mathrm{EM}}(t)$ can be written as

$$\varphi_{\mathrm{EM}}(t) = A \cos[\omega_c t + \psi(t)] \tag{12.20a}$$

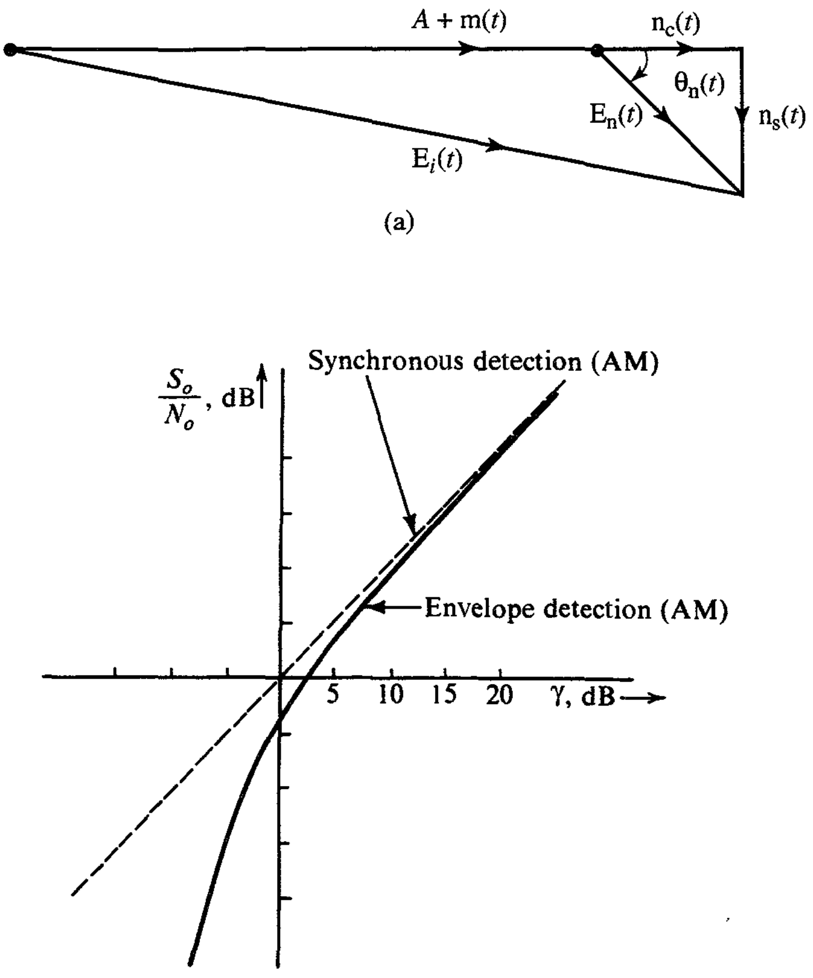
where

$$\psi(t) = k_p \mathrm{m}(t) \qquad \text{for PM} \tag{12.20b}$$

$$= k_f \int_{-\infty}^{t} \mathrm{m}(\alpha)\, d\alpha \qquad \text{for FM} \tag{12.20c}$$

and $\mathrm{m}(t)$ is the message signal. The channel noise $\mathrm{n}(t)$ at the demodulator input is a bandpass noise with PSD $S_\mathrm{n}(\omega)$ and bandwidth $2(\Delta f + B)$. The noise $\mathrm{n}(t)$ can be expressed in terms of quadrature components as

$$\mathrm{n}(t) = \mathrm{n}_c(t) \cos \omega_c t + \mathrm{n}_s(t) \sin \omega_c t \tag{12.21a}$$


A + m(t)
n_c(t)
θ_n(t)
E_n(t)
n_s(t)
E_i(t)
S_o/N_o, dB
Synchronous detection (AM)
Envelope detection (AM)
5
10
15
20
γ, dB

(a)

(b)

Figure 12.5 Performance of AM (synchronous detection and envelope detection).

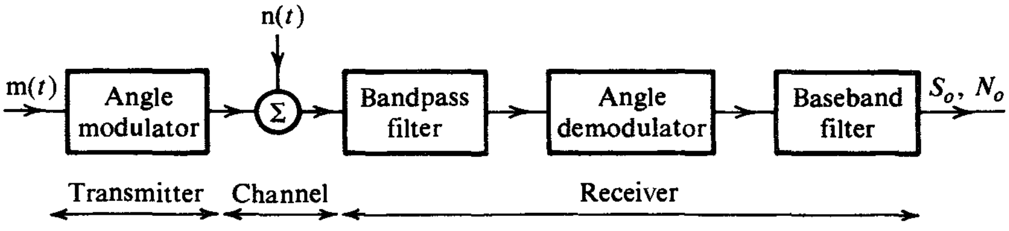

n(t)
m(t)
Angle modulator
Σ
Bandpass filter
Angle demodulator
Baseband filter
S_o, N_o
Transmitter
Channel
Receiver

Figure 12.6 Angle-modulated system.

where $n_c(t)$ and $n_s(t)$ are low-pass signals of bandwidth $\Delta f + B$. The bandpass noise $n(t)$ may also be expressed in terms of the envelope $E_n(t)$ and phase $\Theta_n(t)$ as [see Fig. 12.5a and Eqs. (12.18)]

$$n(t) = E_n(t) \cos [\omega_c t + \Theta_n(t)] \tag{12.21b}$$

Angle modulation (and particularly wide-band angle modulation) is a nonlinear type of modulation. Hence, superposition does not apply. In AM, the signal output can be calculated by assuming the channel noise to be zero, and the noise output can be calculated by assuming the modulating signal to be zero. This is a consequence of linearity. The signal and noise do not form intermodulation components. Unfortunately, exponential modulation is nonlinear, and we cannot use superposition to calculate the output, as can be done in AM. We shall show that because of special circumstances, however, even in angle modulation the noise output can be calculated by assuming the modulating signal to be zero. To prove this we shall first consider the case of PM and then extend those results to FM.

Phase Modulation

Because narrow-band modulation is approximately linear, we need to consider only wide-band modulation. The crux of the argument is that for wide-band modulation, the signal $\mathrm{m}(t)$ changes very slowly relative to the noise $\mathrm{n}(t)$. The modulating signal bandwidth is B, and the noise bandwidth is $2(\Delta f + B)$, with $\Delta f \gg B$. Hence, the phase and frequency variations of the modulated carrier are much slower than are the variations of $\mathrm{n}(t)$. The modulated carrier appears to have constant frequency and phase over several cycles, and, hence, the carrier appears to be unmodulated. We may therefore calculate the output noise by assuming $\mathrm{m}(t)$ to be zero (or a constant). This is a qualitative justification for the linearity argument. A quantitative justification is given in the following development.

To calculate the signal and noise powers at the output, we shall first construct a phasor diagram of the signal $\mathrm{y}_i(t)$ at the demodulator input, as shown in Fig. 12.7,

$$\begin{aligned}\mathrm{y}_i(t) &= A\cos[\omega_c t + \psi(t)] + \mathrm{n}(t)\\ &= A\cos[\omega_c t + \psi(t)] + \mathrm{E_n}(t)\cos[\omega_c t + \Theta_\mathrm{n}(t)]\\ &= \mathrm{R}(t)\cos[\omega_c t + \psi(t) + \Delta\psi(t)]\end{aligned} \tag{12.22}$$

where

$$\psi(t) = k_p\mathrm{m}(t) \qquad \text{for PM} \tag{12.23}$$

For the small-noise case, where $\mathrm{E_n}(t) \ll A$ for "almost all t," $\Delta\psi(t) \ll \pi/2$ for "almost all t," and

$$\Delta\psi(t) \simeq \frac{\mathrm{E_n}(t)}{A}\sin[\Theta_\mathrm{n}(t) - \psi(t)] \tag{12.24}$$

The demodulator detects the phase of the input $\mathrm{y}_i(t)$. Hence, the demodulator output is

$$\mathrm{y}_o(t) = \psi(t) + \Delta\psi(t) \tag{12.25a}$$

$$= k_p\mathrm{m}(t) + \frac{\mathrm{E_n}(t)}{A}\sin[\Theta_\mathrm{n}(t) - \psi(t)] \tag{12.25b}$$

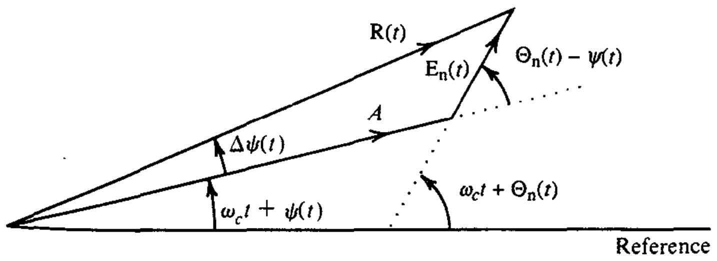


Figure 12.7 Phasor representation of signals in an angle-modulated system.

Note that the noise term $\Delta\psi(t)$ involves the signal $\psi(t)$ due to the nonlinear nature of angle modulation. Because $\psi(t)$ (baseband signal) varies much more slowly than $\Theta_n(t)$ (wide-band noise), we can approximate $\psi(t)$ by a constant ψ,

$$\begin{aligned}\Delta\psi(t) &\simeq \frac{E_n(t)}{A}\sin[\Theta_n(t)-\psi]\\ &= \frac{E_n(t)}{A}\sin\Theta_n(t)\cos\psi - \frac{E_n(t)}{A}\cos\Theta_n(t)\sin\psi\\ &= \frac{n_s(t)}{A}\cos\psi - \frac{n_c(t)}{A}\sin\psi\end{aligned}$$

Also, because $n_c(t)$ and $n_s(t)$ are incoherent for white noise,

$$\begin{aligned}S_{\Delta\psi}(\omega) &= \frac{\cos^2\psi}{A^2}S_{n_s}(\omega) + \frac{\sin^2\psi}{A^2}S_{n_c}(\omega)\\ &= \frac{S_{n_s}(\omega)}{A^2} \qquad [\text{because } S_{n_c}(\omega) = S_{n_s}(\omega)]\end{aligned} \tag{12.26}$$

For a white channel noise with PSD $\mathcal{N}/2$ [Eq. (11.49)],

$$S_{\Delta\psi}(\omega) = \begin{cases}\dfrac{\mathcal{N}}{A^2} & |f| \le \Delta f + B\\ 0 & \text{otherwise}\end{cases} \tag{12.27a}$$

Note that if we had assumed $\psi(t) = 0$ (zero message signal) in Eq. (12.26), we would have obtained exactly the same result.*

The demodulated noise bandwidth is $\Delta f + B$. But because the useful signal bandwidth is only B, the demodulator output is passed through a low-pass filter of bandwidth B to remove the out-of-band noise. Hence, the PSD of the low-pass filter output noise is

$$S_{n_o}(\omega) = \begin{cases}\dfrac{\mathcal{N}}{A^2} & |\omega| \le 2\pi B\\ 0 & |\omega| > 2\pi B\end{cases} \tag{12.27b}$$

and

$$N_o = 2B\left(\frac{\mathcal{N}}{A^2}\right) = \frac{2\mathcal{N}B}{A^2} \tag{12.28a}$$

From Eq. (12.25b) we have

$$S_o = k_p^2\overline{m^2} \tag{12.28b}$$

Thus,

$$\frac{S_o}{N_o} = (Ak_p)^2\frac{\overline{m^2}}{2\mathcal{N}B} \tag{12.28c}$$

These results are valid for small noise, and they apply to both WBPM and NBPM. We also have

$$\gamma = \frac{S_i}{\mathcal{N}B} = \frac{A^2/2}{\mathcal{N}B} = \frac{A^2}{2\mathcal{N}B}$$

---

* This follows from the fact that $E_n(t)\sin\Theta_n(t) = n_s(t)$. Hence, $\Delta\psi(t) = n_s(t)/A$, and $S_{\Delta\psi}(\omega) \simeq S_{n_s}(\omega)/A^2$.

and

$$\frac{S_o}{N_o} = k_p^2 \overline{m^2} \gamma \tag{12.29}$$

Also, for PM [Eq. (5.17a)],

$$\Delta\omega = k_p m_p' \qquad\qquad m_p' = [\dot{m}(t)]_{\max}$$

Hence,

$$\frac{S_o}{N_o} = (\Delta\omega)^2 \left(\frac{\overline{m^2}}{m_p'^2}\right) \gamma \tag{12.30}$$

Note that the bandwidth of the angle-modulated waveform is about $2\Delta f$ (for the wide-band case). Thus, the output SNR increases with the square of the transmission bandwidth; that is, the output SNR increases by 6 dB for each doubling of the transmission bandwidth. Remember, however, that this result is valid only when the noise power is much smaller than the carrier power. Hence, the output SNR cannot be increased indefinitely by increasing the transmission bandwidth because this also increases the noise power, and at some stage the small-noise assumption is violated. When the noise power becomes comparable to the carrier power, the threshold appears as explained later, and a further increase in bandwidth actually reduces the output SNR instead of increasing it.

Let us apply Eq. (12.30) to tone modulation, where $m(t) = \alpha \cos \omega_m t$. For this case $\overline{m^2} = \alpha^2/2$, and $m_p' = |\dot{m}(t)|_{\max} = \alpha\omega_m$. Hence,

$$\frac{S_o}{N_o} = \frac{1}{2}\left(\frac{\Delta\omega}{\omega_m}\right)^2 \gamma = \frac{1}{2}\left(\frac{\Delta f}{f_m}\right)^2 \gamma \tag{12.31}$$

Note that Eqs. (12.28c), (12.29), (12.30), and (12.31) are valid for both NBPM and WBPM.

Frequency Modulation

Frequency modulation may be considered as a special case of phase modulation, where the modulating signal is $\int_{-\infty}^{t} m(\alpha)\, d\alpha$ (Fig. 12.8). At the receiver, we can demodulate FM with a PM demodulator followed by a differentiator, as shown in Fig. 12.8. The PM demodulator output is $k_f \int^t m(\alpha)\, d\alpha$. The subsequent differentiator yields the output $k_f m(t)$, so that

$$S_o = k_f^2 \overline{m^2} \tag{12.32}$$

The phase demodulator output noise will be identical to that calculated earlier, with PSD $\mathcal{N}/A^2$ for white channel noise. This noise is passed through an ideal differentiator whose

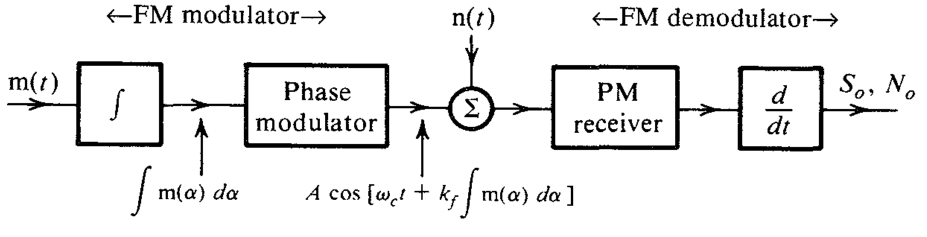


Figure 12.8 FM system as a special case of PM system.

transfer function is $j\omega$. Hence, the PSD $S_{n_o}(\omega)$ of the output noise is $|j\omega|^2$ times the PSD in Eq. (12.27b),

$$S_{n_o}(\omega) = \begin{cases} \dfrac{\mathcal{N}}{A^2}\omega^2 & |\omega| \leq 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \tag{12.33}$$

The PSD of the output noise is parabolic (Fig. 12.9), and the output noise power is

$$\begin{aligned} N_o &= 2\int_0^B \frac{\mathcal{N}}{A^2}(2\pi f)^2\, df \\ &= \frac{8\pi^2 \mathcal{N} B^3}{3A^2} \end{aligned} \tag{12.34}$$

Hence, the output SNR is

$$\begin{aligned} \frac{S_o}{N_o} &= 3\left(\frac{k_f^2\, \overline{m^2}}{(2\pi B)^2}\right)\left(\frac{A^2/2}{\mathcal{N}B}\right) \\ &= 3\left[\frac{k_f^2\, \overline{m^2}}{(2\pi B)^2}\right]\gamma \end{aligned} \tag{12.35}$$

Because $\Delta\omega = k_f m_p$,

$$\frac{S_o}{N_o} = 3\left(\frac{\Delta f}{B}\right)^2\left(\frac{\overline{m^2}}{m_p^2}\right)\gamma \tag{12.36}$$

$$= 3\beta^2\gamma\left(\frac{\overline{m^2}}{m_p^2}\right) \tag{12.37}$$

Recall that the transmission bandwidth is about $2\Delta f$. Hence, for each doubling of the bandwidth, the output SNR increases by 6 dB. Just as in the case of PM, the output SNR does not increase indefinitely because threshold appears as the increased bandwidth makes the channel noise power comparable to the carrier power.

For tone modulation, $\overline{m^2}/m_p^2 = 0.5$ and

$$\frac{S_o}{N_o} = \frac{3}{2}\beta^2\gamma \tag{12.38}$$

The output SNR S_o/N_o (in decibel) in Eq. (12.38) is plotted in Fig. 12.10 as a function of γ (in decibels) for various values of β. The dotted portion of the curves indicates the threshold region (to be discussed later in this section). Although the curves in Fig. 12.10 are valid for tone modulation only ($\overline{m^2}/m_p^2 = 0.5$), they can be used for any other modulating signal m(t) simply by shifting them vertically by a factor $(\overline{m^2}/m_p^2)/0.5 = 2\overline{m^2}/m_p^2$.

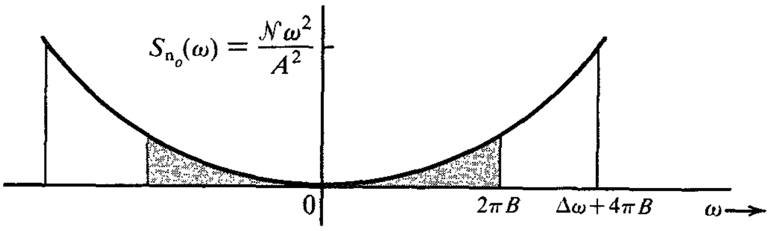


Figure 12.9 PSD of output noise in FM receiver.

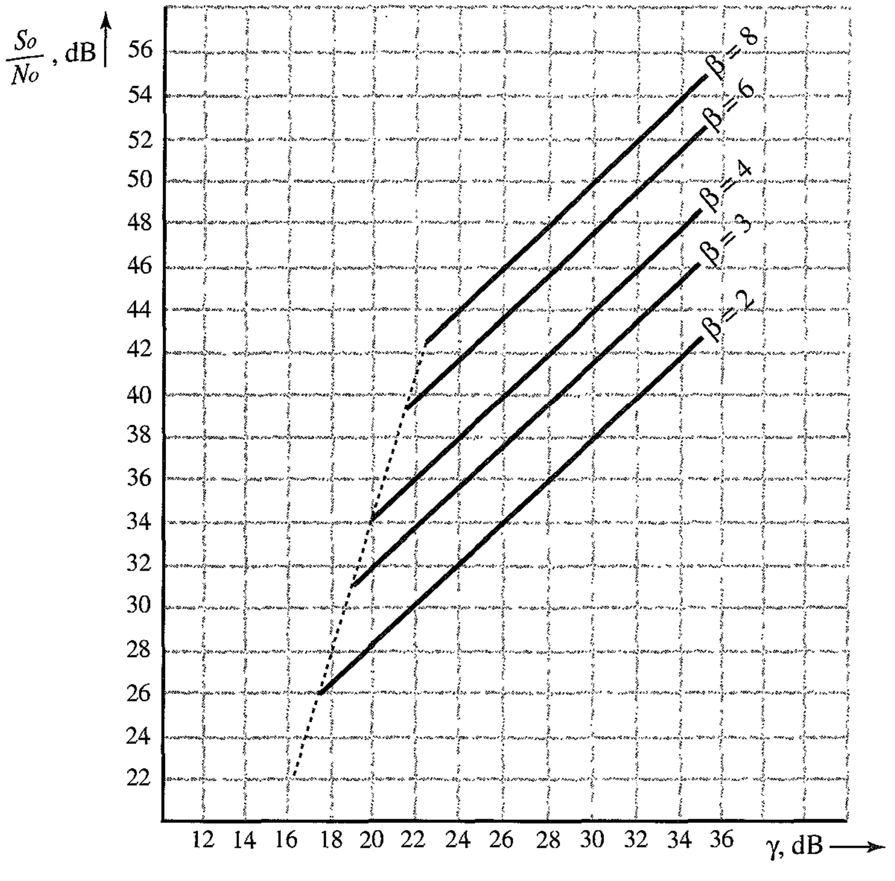


Figure 12.10 Performance of FM system.

From Eqs. (12.31) and (12.38), we observe that for tone modulation FM is superior to PM by a factor of 3. This does not mean that FM is superior to PM for other modulating signals as well. In fact, we shall see that PM is superior to FM for most of the practical signals. From Eqs. (12.30) and (12.36), it can be seen that

$$\frac{(S_o/N_o)_{\text{PM}}}{(S_o/N_o)_{\text{FM}}} = \frac{(2\pi B)^2 m_p^2}{3m_p'^2} \tag{12.39}$$

Hence, if $(2\pi B)^2 m_p^2 > 3m_p'^2$, PM is superior to FM. If the PSD of $\mathrm{m}(t)$ is concentrated at lower frequencies, low-frequency components predominate in $\mathrm{m}(t)$, and m_p' is small. This favors PM. Therefore, in general, PM is superior to FM when $S_{\mathrm{m}}(\omega)$ is concentrated at lower frequencies and FM is superior to PM when $S_{\mathrm{m}}(\omega)$ is concentrated at higher frequencies. This explains why FM is superior to PM for tone modulation, where all the signal power is concentrated at the highest frequency in the band. But for most of the practical signals, the signal power is concentrated at lower frequencies, and PM proves superior to FM.

EXAMPLE 12.3 For a zero mean gaussian random process $\mathrm{m}(t)$ as the baseband signal, determine the output SNR for FM, assuming white gaussian channel noise.

For a gaussian $\mathrm{m}(t)$, $m_p = \infty$. But because the probability that amplitude m lies beyond $3\sigma_{\mathrm{m}}$ ($|\mathrm{m}| \geq 3\sigma_{\mathrm{m}}$) is about 0.0027, one may consider* m_p to be $3\sigma_{\mathrm{m}}$. Hence,

---

* This is known as 3σ **loading**. In the literature, 4σ loading $[m_p = 4\sigma_{\mathrm{m}})]$ is also used.

$$\overline{\mathrm{m}^2} = \sigma_{\mathrm{m}}^2 \qquad \text{and} \qquad m_p^2 = (3\sigma_{\mathrm{m}})^2$$

From Eq. (12.37) it follows that

$$\frac{S_o}{N_o} = \frac{1}{3}\beta^2\gamma$$

Narrow-Band Modulation: The equations derived thus far are valid for both narrow-band and wide-band modulation. We observed in Chapter 5 that narrow-band exponential modulation (NBEM) is approximately linear and is very similar to AM. In fact, the output SNRs for NBEM and AM are similar. To see this, consider the cases of NBPM [Eq. (5.10)] and AM [Eq. (4.8a)],

$$\varphi_{\mathrm{AM}}(t) = A \cos \omega_c t + \mathrm{m}(t) \cos \omega_c t$$

$$\begin{aligned}\varphi_{\mathrm{NBPM}}(t) &= A \cos \omega_c t - Ak_p\mathrm{m}(t) \sin \omega_c t \\ &= A \cos \omega_c t - \mathrm{m}_1(t) \sin \omega_c t\end{aligned}$$

where $\mathrm{m}_1(t) = Ak_p\mathrm{m}(t)$. Both φ_{AM} and φ_{NBPM} contain a carrier and a DSB term. In φ_{NBPM} the carrier and the DSB component are out of phase by $\pi/2$ rad, whereas in φ_{AM} they are in phase. But the $\pi/2$-rad phase difference has no effect on the power. Thus, $\mathrm{m}(t)$ in φ_{AM} is analogous to $\mathrm{m}_1(t)$ in φ_{NBPM}.

Now let us compare the output SNRs for AM and NBPM. For AM [Eq. (12.17)],

$$\left(\frac{S_o}{N_o}\right)_{\mathrm{AM}} = \frac{\overline{\mathrm{m}^2}}{A^2 + \overline{\mathrm{m}^2}}\gamma$$

whereas for NBPM with $\mathrm{m}_1(t) = Ak_p\mathrm{m}(t)$, Eq. (12.29) can be expressed as

$$\begin{aligned}\left(\frac{S_o}{N_o}\right)_{\mathrm{PM}} &= k_p^2\,\overline{\mathrm{m}^2}\,\gamma \\ &= \frac{\overline{\mathrm{m}_1^2}}{A^2}\gamma\end{aligned}$$

Note that for NBPM, we require that $|k_p\mathrm{m}(t)| \ll 1$, that is, $\mathrm{m}_1(t)/A \ll 1$. Hence,

$$A^2 \simeq A^2 + \overline{\mathrm{m}_1^2}$$

and

$$\left(\frac{S_o}{N_o}\right)_{\mathrm{PM}} \simeq \frac{\overline{\mathrm{m}_1^2}}{A^2 + \overline{\mathrm{m}_1^2}}\gamma$$

which is of the same form as $(S_o/N_o)_{\mathrm{AM}}$. Hence, NBPM is very similar to AM. Under the best possible conditions, however, AM outperforms NBPM because for AM, we need only to satisfy the conditions $[A + \mathrm{m}(t)] > 0$ [Eq. (4.9a)], which implies $[\mathrm{m}(t)]_{\max} \le A$. Thus for tone modulation, the modulation index for AM can be nearly equal to unity. For NBPM, however, the narrow-band condition would be equivalent to requiring $\mu \ll 1$. Hence, although AM and NBPM have identical performance for a given value of μ, AM has the edge over NBPM from the SNR viewpoint.

It is interesting to look for the line (in terms of Δf) that separates narrow-band and wide-band FM. We may consider the dividing line to be that value of Δf for which the output SNR for FM given in Eq. (12.37) is equal to the maximum output SNR for AM. The maximum SNR for AM occurs when $\mu = 1$, or when $A = m_p$. Hence, equating Eq. (12.37) with Eq. (12.17) [with $A = m_p$],

$$3\beta^2\gamma\left(\frac{\overline{\mathrm{m}^2}}{m_p^2}\right) = \frac{\overline{\mathrm{m}^2}}{m_p^2 + \overline{\mathrm{m}^2}}\gamma$$

or

$$\beta^2 = \frac{1}{3}\left[\frac{1}{1 + (\overline{\mathrm{m}^2}/m_p^2)}\right] \tag{12.40}$$

Because $\overline{\mathrm{m}^2}/m_p^2 < 1$ for practical signals, and

$$\beta^2 < \frac{1}{3}$$

or

$$\beta < 0.6 \tag{12.41a}$$

This gives

$$\Delta f = 0.6B \tag{12.41b}$$

Mean Square Bandwidth and Estimation of Angle-Modulation Bandwidth

The bandwidth of a signal is a measure of the width of the signal spectrum. Several definitions of bandwidth have appeared in the literature. All are meaningful and useful in different situations. A 3-dB bandwidth is commonly used in electronic circuits.

We now introduce one such useful measure of bandwidth. In Chapter 10, we noted that the standard deviation σ is a good measure of the width of a PDF. We can extend this idea to the PSD by normalizing the spectrum to a unit area. The variance of such a normalized spectrum is known as the **mean square bandwidth** of the spectrum. The mean square bandwidth is meaningful, extremely useful, and mathematically tractable for angle-modulated signals. A normalized PSD, by definition, has unit area. Hence, to normalize a PSD, we divide it by its area so that the resulting function has a unit area. For convenience, we shall use the variable f (in hertz) rather than the variable ω (in radians per second). This will give us the mean square bandwidth in hertz rather than in radians per second.

Using this concept, for a baseband signal $\mathrm{m}(t)$, the normalized PSD is $S_{\mathrm{m}}(\omega)/\int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)\, df$. Because the normalized PSD is symmetrical about the vertical axis ($f = 0$), it has a zero mean (in the sense of the PDF), and its variance $\overline{B_{\mathrm{m}}^2}$ is*

$$\overline{B_{\mathrm{m}}^2} = \frac{\int_{-\infty}^{\infty} f^2 S_{\mathrm{m}}(2\pi f)\, df}{\int_{-\infty}^{\infty} S_{\mathrm{m}}(2\pi f)\, df} \text{ Hz} \tag{12.42a}$$

$$= \frac{1}{\overline{\mathrm{m}^2}} \int_{-\infty}^{\infty} f^2 S_{\mathrm{m}}(2\pi f)\, df \text{ Hz} \tag{12.42b}$$

---

* For $\overline{B_{\mathrm{m}}^2}$ to exist, $S_{\mathrm{m}}(\omega)$ must approach zero at a rate faster than $1/\omega^2$ for large values of ω.

EXAMPLE 12.4 For a low-pass signal with PSD $S_m(\omega) = \text{rect}(\omega/4\pi B)$, show that $\overline{B_m^2} = B^2/3$.

Because $S_m(\omega) = 1$ for $|f| < B$, and 0 for $|f| > B$, we have [Eqs. (12.42)]

$$\overline{B_m^2} = \frac{\int_{-B}^{B} f^2\,df}{\int_{-B}^{B} df} = \frac{B^2}{3}$$

EXAMPLE 12.5 For a gaussian PSD $S_m(\omega) = ke^{-\omega^2/2\sigma^2}$ show that

$$\overline{B_m^2} = \left(\frac{\sigma}{2\pi}\right)^2$$

From Eqs. (12.42),

$$\overline{B_m^2} = \frac{k\int_{-\infty}^{\infty} f^2 e^{-4\pi^2 f^2/2\sigma^2}\,df}{k\int_{-\infty}^{\infty} e^{-4\pi^2 f^2/2\sigma^2}\,df} = \frac{\sigma^2}{4\pi^2}$$

We shall now investigate the dependence of the FM wave PSD on instantaneous frequency. In wideband modulation instantaneous frequency f_i in the range $(f,\ f+df)$ gives rise to power spectral components (components of the PSD) in the range $(f,\ f+df)$. The power contribution of these components is proportional to the relative time that f_i remains in this range. If $p_{f_i}(f)$ is the PDF of instantaneous frequency f_i, then $p_{f_i}(f)\,df$ is the probability of observing f_i in the range $(f,\ f+df)$. This probability is proportional to the time that f_i remains in the range $(f,\ f+df)$ and, hence, is proportional to the power contributed by spectral components in the range $(f,\ f+df)$. If $S_{FM}(\omega)$ is the PSD of an FM wave, then the preceding argument implies*

$$S_{FM}(2\pi f) = kp_{f_i}(f) \qquad f > 0 \tag{12.43}$$

where k is a constant of proportionality.†

To find the mean square bandwidth of $S_{FM}(2\pi f)$, we observe from Eq. (12.43) that $p_{f_i}(f)$ is precisely $S_{FM}(2\pi f)$ normalized to unit area. Hence, the mean square bandwidth of $S_{FM}(2\pi f)$ is the variance $\sigma_{f_i}{}^2$ of f_i, because

$$f_i = \frac{\omega_i}{2\pi} = \frac{1}{2\pi}[\omega_c + k_f m(t)]$$

$$\overline{f_i} = \frac{1}{2\pi}[\omega_c + k_f\overline{m(t)}]$$

$$= f_c \qquad [\text{because } \overline{m(t)} = 0]$$

---

* Because $p_{f_i}(f) = 0$ for $f < 0$, $S_{FM}(\omega)$ in Eq. (12.43) is the unilateral PSD. This means $S_{FM}(\omega) = 0$ for $\omega < 0$ and is twice the bilateral PSD for $\omega > 0$.

† By integrating both sides of Eq. (12.43) over $(0,\ \infty)$, it can be shown that $k = A^2/2$.

and

$$\begin{aligned}\sigma_{\mathrm{f}_i}^2 &= \overline{(\mathrm{f}_i - f_c)^2} \\ &= \frac{1}{4\pi^2}\overline{[k_f \mathrm{m}(t)]^2} \\ &= \frac{1}{4\pi^2} k_f^2 \overline{\mathrm{m}^2}\end{aligned}$$

In other words,

$$\overline{B_{\mathrm{FM}}^2} = \frac{1}{4\pi^2} k_f^2 \overline{\mathrm{m}^2} \tag{12.44}$$

For phase modulation,

$$\mathrm{f}_i = \frac{1}{2\pi}\omega_i = \frac{1}{2\pi}[\omega_c + k_p \dot{\mathrm{m}}(t)]$$

and

$$\begin{aligned}\sigma_{\mathrm{f}_i}^2 &= \frac{1}{(2\pi)^2}\overline{(\mathrm{f}_i - f_c)^2} \\ &= \frac{1}{4\pi^2} k_p^2 \overline{[\dot{\mathrm{m}}(t)]^2}\end{aligned}$$

Because the PSD of $\dot{\mathrm{m}}(t)$ is $\omega^2 S_{\mathrm{m}}(\omega) = 4\pi^2 f^2 S_{\mathrm{m}}(2\pi f)$,

$$\sigma_{\mathrm{f}_i}^2 = \frac{1}{4\pi^2} k_p^2 \int_{-\infty}^{\infty} 4\pi^2 f^2 S_{\mathrm{m}}(2\pi f)\, df$$

Using Eq. (12.42),

$$\sigma_{\mathrm{f}_i}^2 = k_p^2 \overline{\mathrm{m}^2}\, \overline{B_{\mathrm{m}}^2}$$

that is,

$$\overline{B_{\mathrm{PM}}^2} = k_p^2 \overline{\mathrm{m}^2}\, \overline{B_{\mathrm{m}}^2} \tag{12.45}$$

From Eqs. (12.44) and (12.45) we observe that the bandwidth of the FM wave is independent of the modulating signal spectrum, whereas the bandwidth of the PM wave is strongly influenced by the modulating signal spectrum.

The output SNRs in Eqs. (12.29) and (12.35) can now be expressed in terms of mean square bandwidths. For PM,

$$\frac{S_o}{N_o} = \frac{\overline{B_{\mathrm{PM}}^2}}{\overline{B_{\mathrm{m}}^2}}\gamma \tag{12.46}$$

and for FM,

$$\frac{S_o}{N_o} = 3\frac{\overline{B_{\mathrm{FM}}^2}}{B^2}\gamma \tag{12.47}$$

Quantities $\overline{B_{\mathrm{PM}}^2}$ and $\overline{B_{\mathrm{FM}}^2}$ are the variances of the normalized PSDs of the PM and FM waves, respectively. As seen earlier, the actual transmission bandwidth will be several times the standard deviation $\sqrt{\overline{B^2_{\text{PM}}}}$ or $\sqrt{\overline{B^2_{\text{FM}}}}$. For example, when the modulated signal PSD has a gaussian form,* 99.74% of the total power resides within 3σ of the carrier frequency. Hence, the bandwidth in this case may be taken as 6σ, or $6\sqrt{\overline{B^2_{\text{PM}}}}$ for PM and $6\sqrt{\overline{B^2_{\text{FM}}}}$ for FM. From Eqs. (12.46) and (12.47), it follows that in PM as well as in FM the output SNR improves by 6 dB for each doubling of the transmission bandwidth. It should be stressed that these results are valid only for small noise.

Which Is Superior: PM or FM? From Eqs. (12.46) and (12.47), we have

$$\frac{(S_o/N_o)_{\text{PM}}}{(S_o/N_o)_{\text{FM}}} = \left(\frac{B^2}{3\overline{B^2_{\text{m}}}}\right)\left(\frac{\overline{B^2_{\text{PM}}}}{\overline{B^2_{\text{FM}}}}\right) \tag{12.48}$$

It will be instructive to compare the performance of PM and FM for the same transmission bandwidth (the same mean square bandwidth), that is, for $\overline{B^2_{\text{PM}}} = \overline{B^2_{\text{FM}}}$. This gives

$$\frac{(S_o/N_o)_{\text{PM}}}{(S_o/N_o)_{\text{FM}}} = \frac{B^2}{3\overline{B^2_{\text{m}}}} \tag{12.49a}$$

Thus, if

$$B^2 > 3\overline{B^2_{\text{m}}} \tag{12.49b}$$

PM is superior to FM. Otherwise, FM is superior to PM. We showed in Example 12.4 that when the PSD of m(t) is constant (over a band B Hz), $3\overline{B^2_{\text{m}}} = B^2$. Hence, PM and FM perform equally well in that case. If the spectrum falls off with frequency, as it does for all real signals, $\overline{B^2_{\text{m}}}$ is less than $B^2/3$, and PM is superior to FM. If, on the other hand, the spectrum is weighted heavily at higher frequencies, $\overline{B^2_m}$ is greater than $B^2/3$, and FM is superior to PM. This is exactly what happens for tone modulation, where the spectrum is concentrated at the highest frequency B, with no power at lower frequencies, making $\overline{B^2_{\text{m}}} = B^2$. This is why for tone modulation FM is superior to PM by a factor of 3. Tone modulation proves to be grossly misleading in the SNR analysis of angle modulation. For practical signals, including audio, the spectrum falls off with frequency, and PM is superior to FM. Actually, so-called FM broadcast is not pure FM but is FM modified by preemphasis-deemphasis, as discussed in Sec. 5.5.

We have derived two different criteria [Eqs. (12.39) and (12.49a)] for comparing PM and FM performance. In Eq. (12.39), the output SNRs are compared for a given transmission bandwidth, whereas in Eq. (12.49a), the output SNRs are compared for a given mean square transmission bandwidth. Consequently, we may get slightly different answers by using these two criteria (see Probs. 12.3-4 and 12.3-8). Although the criterion in Eq. (12.39) is preferable to that in Eq. (12.49a), it is impossible in practice to determine the parameters required in Eq. (12.39), and the only way to compare may be through Eq. (12.49a).

EXAMPLE 12.6 If a baseband signal m(t) has a gaussian PSD, show that PM is superior to FM by a factor of 3 (4.77 dB) when the bandwidth B is taken as 3σ, where σ is the standard deviation of the normalized PSD of m(t).

---

* It can be shown[2] that for wide-band angle modulation, the PSD of the modulated carrier is gaussian when the modulating random process is gaussian.

$S_m(t)$ is gaussian of the form $ke^{-\omega^2/2\sigma^2}$. The radian bandwidth $W = 2\pi B = 3\sigma$, and $B = 3\sigma/2\pi$. Also, as seen in Example 12.5,

$$\overline{B_m^2} = \frac{\sigma^2}{4\pi^2} \quad \text{and} \quad B = \frac{3\sigma}{2\pi}$$

Hence, from Eqs. (12.49), it follows that PM is superior to FM by a factor of 3.

Threshold in Angle Modulation

When the noise power at the demodulator input is comparable to the carrier power, the threshold phenomenon (discussed earlier for AM systems) appears. In FM this effect is much more pronounced than in AM. Let us discuss qualitatively how the threshold effect appears. We refer back to Fig. 12.7 and observe that the phasor E_n rotates from the terminal of the phasor A. When $E_n \ll A$, the angle $\Delta\psi(t)$ is quite small, and because $\Theta_n(t)$ is random with uniform distribution in the range $(0, 2\pi)$, $\Delta\psi(t)$ assumes positive as well as negative values (Fig. 12.11a), which are usually much smaller than 2π. When E_n is large (on the order of A or greater), however, the resultant phasor is much more likely to rotate around the origin, and $\Delta\psi$ is more likely to go through changes of 2π (Fig. 12.11c) in a relatively short time, because the noise varies much faster than the modulating signal. The noise at the FM demodulator output is given by $\Delta\dot{\psi}(t)$.* This is shown in Fig. 12.11 for large and small noise. For large noise, we observe the appearance of spikes (of area 2π), which give rise to a crackling sound. When the noise is small ($E_n \ll A$), the PSD of the output noise $\Delta\dot{\psi}(t)$ is parabolic, and most of its power is in the frequencies greater than B and is therefore filtered out by the baseband filter at the output. For the large-noise case, on the other hand, we have the presence of spikes, which are like

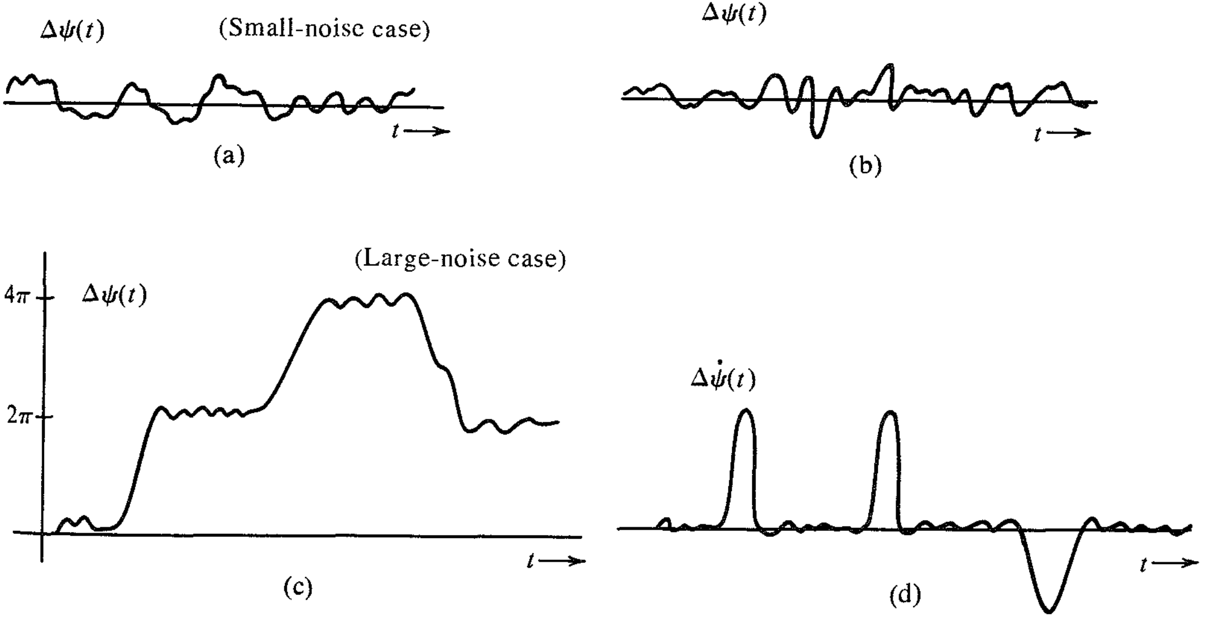


Figure 12.11 Nature of output noise in FM receiver for small and large channel noise.

---

* Recall that FM demodulation requires differentiation of the angle because the instantaneous frequency is proportional to the derivative of the angle.

impulses. Consequently, they have considerable power at lower frequencies. Hence, a spike will contribute much more noise at the output. For this reason, when E_n approaches the order of A, the output noise starts increasing disproportionately (Fig. 12.12). This is precisely the phenomenon of threshold.

It has been shown[3] that the noise power caused by the spikes N_s, is

$$N_s = \frac{8\pi^2 B_{\text{FM}} B}{\sqrt{3}} Q\left(\sqrt{\frac{2B}{B_{\text{FM}}}\gamma}\right) \tag{12.50a}$$

and the total noise power N_T, is the sum of N_o in Eq. (12.34) and N_s in Eq. (12.50a). The output SNR is

$$\begin{aligned}\frac{S_o}{N_o + N_s} &= \frac{3\beta^2\gamma(\overline{\text{m}^2}/m_p^2)}{1 + (2\sqrt{3}B_{\text{FM}}/B)\,\gamma\, Q\left[\sqrt{(2B/B_{\text{FM}})\gamma}\right]} \\ &= \frac{3\beta^2\gamma(\overline{\text{m}^2}/m_p^2)}{1 + 4\sqrt{3}(\beta+1)\gamma Q[\sqrt{\gamma/(\beta+1)}]}\end{aligned} \tag{12.50b}$$

The onset of the threshold is when the carrier power is 10 times the channel noise power. The carrier power is $A^2/2$, and for white noise with a PSD $\mathcal{N}/2$, the noise power is $\mathcal{N}B_{\text{FM}}$, where B_{FM}, the bandwidth of an FM carrier, is

$$B_{\text{FM}} = 2(\Delta f + B) = 2B(\beta + 1)$$

Hence, the threshold occurs when

$$2\mathcal{N}B(\beta + 1) = \frac{1}{10}\frac{A^2}{2}$$

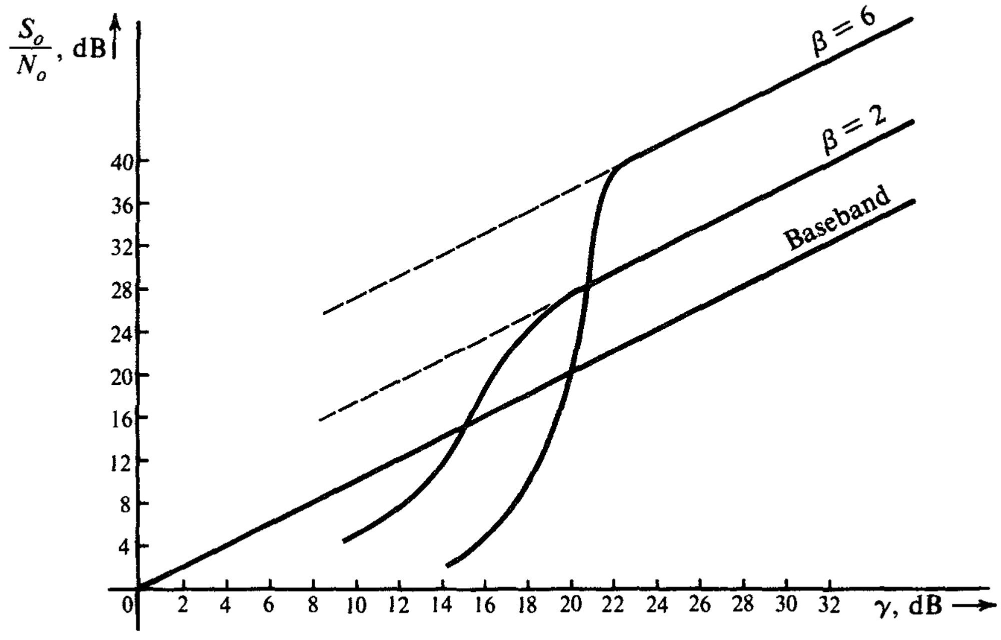


Figure 12.12 Threshold phenomenon in FM.

or

$$20(\beta + 1) = \frac{A^2}{2\mathcal{N}B}$$

By definition, the right-hand side of this equation is γ at threshold. Thus, γ_{thresh}, the value of γ at the onset of threshold, is

$$\gamma_{\text{thresh}} = 20(\beta + 1) \tag{12.51}$$

EXAMPLE 12.7 A gaussian $m(t)$ with 4σ loading (that is, $m_p = 4\sigma_m$) frequency-modulates a carrier using $\beta = 4$. The output SNR is found to be 20.5 dB. Determine whether the system is in threshold.

For $\beta = 4$, γ_{thresh} is

$$\gamma_{\text{thresh}} = 20(\beta + 1) = 20(5) = 100$$

For this value of γ [Eq. (12.37)],

$$\left(\frac{S_o}{N_o}\right) = 3(16)(100)\left(\frac{1}{16}\right) = 300$$

Because 20.5 dB is a ratio of 160, the SNR is below 300, and the system is in threshold.

Threshold Extension in Angle Modulation

The problem of threshold is rather serious in angle modulation. The ability to communicate even at low power levels simply by increasing the transmission bandwidth is the very raison d'être of angle-modulated systems, and threshold deprives angle modulation of its very essence. For this reason, much attention has been paid to the problem of pushing the threshold level further down and extending the useful operating range. The two principal methods used for this purpose utilize frequency compression feedback (FCF) and phase-lock loop (PLL). We shall look briefly into these techniques.

Frequency Compression Feedback (FCF): The basic circuit is shown in Fig. 12.13. Let us consider a frequency-modulated wave $A \cos[\omega_c t + \psi(t)]$ at the input. The quiescent frequency of the VCO is $\omega_c - \omega_o$. The instantaneous frequency ω_i of the VCO is given by

$$\omega_i = (\omega_c - \omega_o) + \alpha e(t)$$

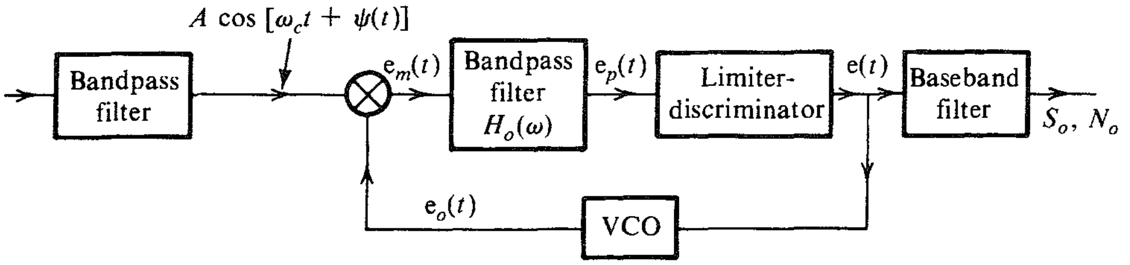


Figure 12.13 Frequency compression feedback (FCF).

and the VCO output $e_o(t)$ is given by* (assuming initial instant $t = 0$),

$$e_o(t) = 2\cos\left[(\omega_c - \omega_o)t + \alpha\int_0^t e(\tau)\,d\tau\right] \tag{12.52}$$

The bandpass filter $H_o(\omega)$ is centered at ω_o. The multiplier output $e_m(t)$ has a spectrum centered at the sum and difference frequencies. The sum frequencies are centered at $2\omega_c - \omega_o$ and are suppressed by the bandpass filter $H_o(\omega)$, which allows only the difference frequencies centered at ω_o to pass. The output $e_p(t)$ of the bandpass filter is

$$e_p(t) = A\cos\left[\omega_o t + \psi(t) - \alpha\int_0^t e(\tau)\,d\tau\right] \tag{12.53}$$

The limiter-discriminator frequency demodulates the signal $e_p(t)$ by differentiating its angle. Therefore, the output $e(t)$ is given by

$$\begin{aligned} e(t) &= \frac{d}{dt}\left[\psi(t) - \alpha\int_0^t e(\tau)\,d\tau\right] \\ &= \dot{\psi}(t) - \alpha e(t) \end{aligned}$$

Hence,

$$e(t) = \frac{\dot{\psi}(t)}{1+\alpha}$$

Substitution of this value of $e(t)$ into Eq. (12.53) yields

$$e_p(t) = A\cos\left[\omega_o t + \frac{\psi(t)}{1+\alpha}\right] \tag{12.54}$$

This interesting result shows that the signal $e_p(t)$ is another frequency-modulated signal with carrier frequency ω_o and is similar to the incoming carrier $A\cos[\omega_c t + \psi(t)]$. The difference is that the angle is $\psi(t)/(1+\alpha)$ instead of $\psi(t)$. This implies a reduction in the frequency deviation and, consequently, a reduction in the bandwidth of the modulated signal by a factor† of $(1+\alpha)$. Hence, for a wide-band case, the bandwidth of $H_o(\omega)$ need only be about $2\Delta f/(1+\alpha)$. The second conclusion is that when $e_p(t)$ is applied to a limiter-discriminator, the output $e(t)$ is $\dot{\psi}(t)/(1+\alpha)$. Hence, the FCF demodulator indeed frequency-demodulates the incoming frequency-modulated carrier.

Let us now see what happens when the input signal is a frequency-modulated carrier plus bandpass channel noise. The signal plus noise can be expressed as $R(t)\cos[\omega_c t + \psi(t) + \Delta\psi(t)]$ [Eq. (12.22)]. The amplitude variations $R(t)$ are eventually eliminated in the limiter-discriminator (see Sec. 5.4), and

$$e_p(t) = A\cos\left[\omega_c t + \frac{\psi(t) + \Delta\psi(t)}{1+\alpha}\right] \tag{12.55}$$

Similarly,

$$e(t) = \frac{1}{1+\alpha}[\dot{\psi}(t) + \Delta\dot{\psi}(t)]$$

---

* The amplitude of $e_o(t)$ is immaterial in our discussion. For convenience, it is considered to be 2.

† In order to avoid distortion of the modulating signal, the bandwidth $(\Delta f + B)/(1+\alpha)$ must be greater than B.

where $\dot{\psi}$ is the useful signal and $\Delta\dot{\psi}$ is the noise. The output of the FCF demodulator is identical to that of the conventional demodulator except for the multiplicative factor $1/(1+\alpha)$. Hence, the output SNR of the FCF demodulator is identical to that of the conventional demodulator. The advantage is gained, however, in extending the threshold. This can be seen from the fact that the FM wave passes unmolested through the filter $H_o(\omega)$ [Eq. (12.55)]. The channel noise, however, which had hitherto a bandwidth of about $2\Delta f$, has to pass through $H_o(\omega)$ having $1/(1+\alpha)$ times its former bandwidth. Hence, the carrier-power-to-noise ratio at the input of the limiter-discriminator is enhanced by a factor of $1+\alpha$. As seen earlier, this does not affect the SNR, but it does extend the threshold because of the relative enhancement of the carrier power in relation to noise power. It can be seen that the carrier power can now afford to be reduced by a factor of $1+\alpha$ before the onset of the threshold. In other words, the threshold is extended roughly by $10 \log \alpha$ dB. Because of practical problems, however, this benefit is not completely realized. In practical FCF demodulators, threshold extension of about 5 to 7 dB can be realized.

Phase-Locked Loop (PLL): The functioning of the PLL was discussed in Chapters 4 and 5. For the small-noise case, the PLL performance is identical to that of a conventional demodulator (just as in FCF). For large noise, however, the PLL extends the threshold just as the FCF does, by reducing the filter bandwidth. The detailed analysis is beyond our scope.[4] A practical PLL extends the threshold region by about 3 to 6 dB.

12.4 PULSE-MODULATED SYSTEMS

Among pulse-modulated systems (PAM, PWM, PPM, and PCM), only PCM is of practical importance. The other systems are rarely used in practice. For this reason we shall discuss only PCM in detail. It can be shown that the performance of PAM is similar to that of AM-SC systems (i.e., the output SNR is equal to γ). The PWM and PPM systems are capable of exchanging the transmission bandwidth with output SNR, as in angle-modulated systems. In PWM, the output SNR is proportional to the transmission bandwidth B_T. This performance is clearly inferior to that of angle-modulated systems, where the output SNR increases as B_T^2. In PPM systems under optimum conditions, the output SNR increases as B_T^2 but is still inferior to FM by a factor of 6. For in-depth treatment of PAM, PWM, and PPM, the reader is referred to Panter[1] or Rowe.[5]

Another PM system that deserves mention is the delta-modulation (DM) system discussed in Chapter 6. For speech signals, this system's performance is comparable to that of PCM for a bandwidth expansion ratio B_T/B of 7 to 8. For $B_T/B > 8$, PCM is superior to DM, and for $B_T/B < 8$, DM is superior to PCM.

Pulse-Code Modulation

In PCM, a baseband signal m(t) band-limited to B Hz and with amplitudes in the range of $-m_p$ to m_p is sampled at a rate of $2B$ samples per second. The sample amplitudes are quantized into L levels, each separated by $2m_p/L$. Each quantized sample is encoded into n binary digits ($2^n = L$). The binary signal is transmitted over a channel. The receiver detects the binary signal and reconstructs quantized samples (decoding). The quantized samples are then passed through a low-pass filter to obtain the desired signal m(t).

There are two sources of error in PCM: (1) quantization or "rounding off" error, and (2) detection error. The latter is caused by error in the detection of the binary signal at the receiver.

As usual, $\mathrm{m}(t)$ is assumed to be a wide-sense stationary random process. The random variable $\mathrm{m}(kT_s)$, formed by sample-function amplitudes at $t = kT_s$, will be denoted by m_k. The kth sample m_k is rounded off, or quantized, to a value $\hat{\mathrm{m}}_k$, which is encoded and transmitted as binary digits. Because of the channel noise, some of the digits may be detected erroneously at the receiver, and the reconstructed sample will be $\tilde{\mathrm{m}}_k$ instead of $\hat{\mathrm{m}}_k$. If q_k and ϵ_k are the quantization and detection errors, respectively, then

$$\mathrm{q}_k = \mathrm{m}_k - \hat{\mathrm{m}}_k \tag{12.56}$$

$$c_k = \hat{\mathrm{m}}_k - \tilde{\mathrm{m}}_k$$

and

$$\mathrm{m}_k - \tilde{\mathrm{m}}_k = \mathrm{q}_k + \epsilon_k \tag{12.57}$$

Hence, the total error $\mathrm{m}_k - \tilde{\mathrm{m}}_k$ at the receiver is $\mathrm{q}_k + \epsilon_k$. The receiver reconstructs the signal $\tilde{\mathrm{m}}(t)$ from samples $\tilde{\mathrm{m}}_k$ according to the interpolation formula in Eq. (6.10),

$$\begin{aligned}
\tilde{\mathrm{m}}(t) &= \sum_k \tilde{\mathrm{m}}_k \operatorname{sinc}(2\pi Bt - k\pi) \\
&= \sum_k [\mathrm{m}_k - (\mathrm{q}_k + \epsilon_k)] \operatorname{sinc}(2\pi Bt - k\pi) \\
&= \sum_k \mathrm{m}_k \operatorname{sinc}(2\pi Bt - k\pi) - \sum_k (q_k + \epsilon_k) \operatorname{sinc}(2\pi Bt - k\pi) \\
&= \mathrm{m}(t) - \mathrm{e}(t)
\end{aligned} \tag{12.58a}$$

where

$$\mathrm{e}(t) = \sum_k (\mathrm{q}_k + \epsilon_k) \operatorname{sinc}(2\pi Bt - k\pi) \tag{12.58b}$$

The receiver therefore receives the signal $\mathrm{m}(t) - \mathrm{e}(t)$ instead of $\mathrm{m}(t)$. The error signal $\mathrm{e}(t)$ is a random process with the kth sample $\mathrm{q}_k + \epsilon_k$. Because the process is wide-sense stationary, the mean square value of the process is the same as the mean square value at any instant. Because $\mathrm{q}_k + \epsilon_k$ is the value of $\mathrm{e}(t)$ at $t = kT_s$,

$$\overline{\mathrm{e}^2(t)} = \overline{(\mathrm{q}_k + \epsilon_k)^2}$$

Because q_k and ϵ_k are independent with zero mean RVs (see Examples 10.17, 10.18, and 10.19),

$$\overline{\mathrm{e}^2(t)} = \overline{\mathrm{q}_k^2} + \overline{\epsilon_k^2}$$

We have already derived $\overline{\mathrm{q}_k^2}$ and $\overline{\epsilon_k^2}$ in Examples 10.17 and 10.18 [Eqs. (10.64b) and (10.66b)]. Hence,

$$\overline{\mathrm{e}^2(t)} = \frac{1}{3}\left(\frac{m_p}{L}\right)^2 + \frac{4m_p^2 P_e(L^2 - 1)}{3L^2} \tag{12.59a}$$

where P_e is the detection error probability. For binary coding, each sample is encoded into n binary digits. Hence, $2^n = L$, and

$$\overline{\mathrm{e}^2(t)} = \frac{m_p^2}{3(2^{2n})}[1 + 4P_e(2^{2n} - 1)] \tag{12.59b}$$

As seen from Eq. (12.58a), the output $\mathrm{m}(t) - \mathrm{e}(t)$ contains the signal $\mathrm{m}(t)$ and noise $\mathrm{e}(t)$. Hence,

$$S_o = \overline{\mathrm{m}^2} \qquad N_o = \overline{\mathrm{e}^2(t)}$$

and

$$\frac{S_o}{N_o} = \frac{3(2^{2n})}{1 + 4P_e(2^{2n} - 1)} \left(\frac{\overline{\mathrm{m}^2}}{m_p^2}\right) \tag{12.60}$$

The error probability P_e depends on A_p, the peak pulse amplitude, and the channel noise power σ_n^2 [Eq. (10.41)],*

$$P_e = Q\left(\frac{A_p}{\sigma_\mathrm{n}}\right)$$

It will be shown in Sec. 13.1 that A_p/σ_n can be maximized (that is, P_e can be minimized) by passing the incoming digital signal through an optimum filter (known as the **matched filter**). It will be shown that for polar signaling [Eqs. (13.9c) and (13.20a)],

$$\left(\frac{A_p}{\sigma_\mathrm{n}}\right)_{\max} = \sqrt{\frac{2E_p}{\mathcal{N}}}$$

and

$$(P_e)_{\min} = Q\left(\sqrt{\frac{2E_p}{\mathcal{N}}}\right) \tag{12.61}$$

where E_p is the energy of the received binary pulse and the channel noise is assumed to be white with PSD $\mathcal{N}/2$. Because there are n binary pulses per sample and there are $2B$ samples per second, there are a total of $2Bn$ pulses per second. Hence, the received signal power $S_i = 2BnE_p$, and

$$\begin{aligned} P_e &= Q\left(\sqrt{\frac{S_i}{n\mathcal{N}B}}\right) \\ &= Q\left(\sqrt{\frac{\gamma}{n}}\right) \end{aligned} \tag{12.62}$$

and

$$\frac{S_o}{N_o} = \frac{3(2^{2n})}{1 + 4(2^{2n} - 1)Q\left(\sqrt{\gamma/n}\right)} \left(\frac{\overline{\mathrm{m}^2}}{m_p^2}\right) \tag{12.63}$$

---

* This assumes polar signaling. Bipolar (or pseudoternary) signaling requires about 3 dB more power than polar to achieve the same P_e (see Sec. 7.2). In practice, bipolar rather than polar signaling is used for PCM.

Figure 12.14 shows a plot of the output SNR as a function of γ for tone modulation ($\overline{m^2}/m_p^2 = 0.5$). It can also be used for a general case, with $\overline{m^2}/m_p^2 = \lambda$, where the SNR is $\lambda/0.5 = 2\lambda$ times that in Fig. 12.14. Hence, we simply shift the curves in Fig. 12.14 upward by $10 \log (2\lambda)$ dB.

Figure 12.14 shows two interesting features: the threshold and the saturation. When γ is too small, a large pulse-detection error results, and the decoded pulse sequence yields a sample value that has no relation to the actual sample transmitted. The received signal is thus meaningless, and we have the phenomenon of threshold. To explain saturation, we observe that when γ is sufficiently large (implying sufficiently large pulse amplitude), the detection error $P_e \to 0$, and Eq. (12.60) becomes

$$\frac{S_o}{N_o} = 3L^2 \left(\frac{\overline{m^2}}{m_p^2}\right) \tag{12.64a}$$

$$= 3(2^{2n}) \left(\frac{\overline{m^2}}{m_p^2}\right) \tag{12.64b}$$

The SNR in this case is practically independent of γ. Because the detection error approaches zero, the output noise now consists entirely of the quantization noise, which depends only on

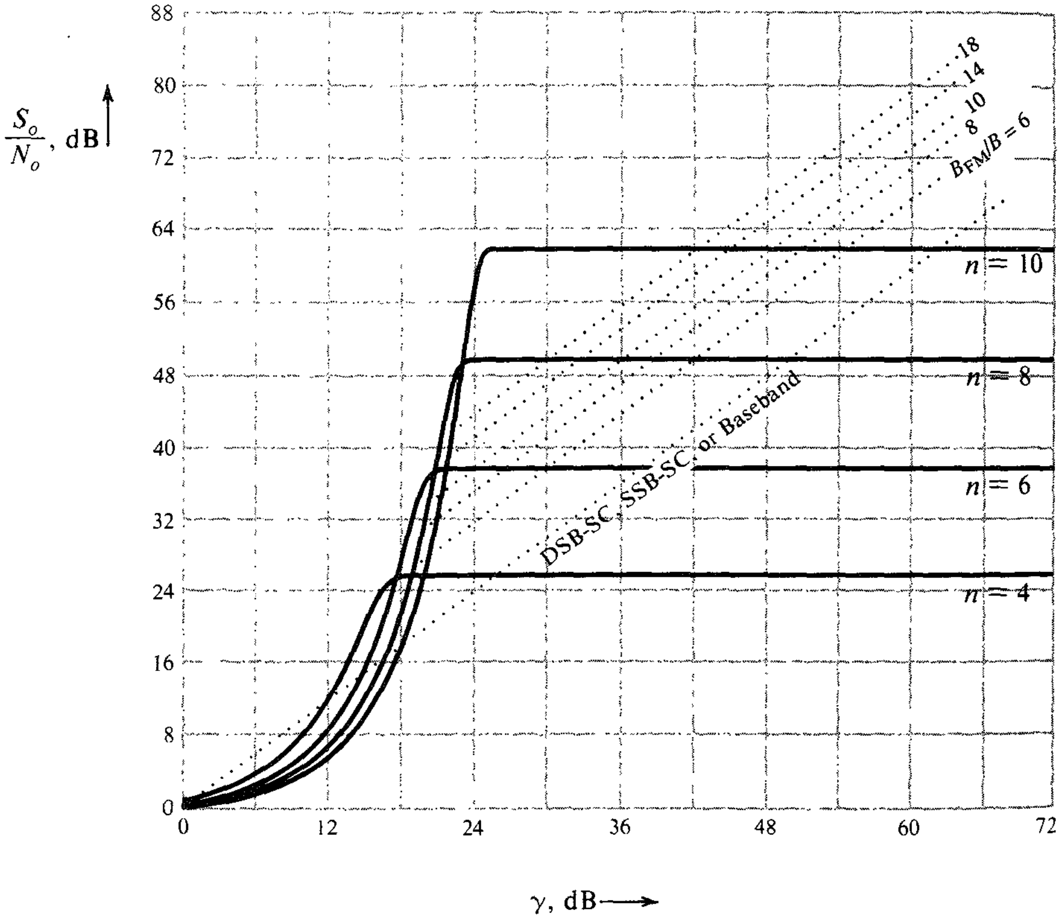


Figure 12.14 Performance of PCM.

L. Because the pulse amplitude is large enough so that there is very little probability of making a detection error, a further increase in γ by increasing the pulse amplitude buys no advantage, and we have the saturation effect.

In the saturation region,

$$\left(\frac{S_o}{N_o}\right)_{\text{dB}} = 10\left[\log 3 + 2n\log 2 + \log\left(\frac{\overline{\mathrm{m}^2}}{m_p^2}\right)\right]$$
$$= \alpha + 6n \tag{12.64c}$$

where $\alpha = 4.77 + 10\log_{10}(\overline{\mathrm{m}^2}/m_p^2)$.

EXAMPLE 12.8 For PCM with $n = 8$, determine the output SNR for a gaussian $\mathrm{m}(t)$. Assume the saturation region of operation.

For a gaussian signal, $m_p = \infty$. In practice, however, we may clip amplitudes $> 3\sigma_{\mathrm{m}}$ or $4\sigma_{\mathrm{m}}$, depending on the accuracy desired. For example, in the case of 3σ loading,

$$P(|\mathrm{m}| > 3\sigma_{\mathrm{m}}) = 2Q(3) = 0.0026$$

and for 4σ loading,

$$P(|\mathrm{m}| > 4\sigma_{\mathrm{m}}) = 2Q(4) = 6\times 10^{-5}$$

If we take the case of 3σ loading,

$$\frac{\overline{\mathrm{m}^2}}{m_p^2} = \frac{\sigma_{\mathrm{m}}^2}{(3\sigma_{\mathrm{m}})^2} = \frac{1}{9}$$

and

$$\frac{S_o}{N_o} = 3(2)^{16}\left(\frac{1}{9}\right) = 21{,}845 = 43.4\text{ dB}$$

For 4σ loading,

$$\frac{S_o}{N_o} = 3(2)^{16}\left(\frac{1}{16}\right) = 12{,}288 = 40.9\text{ dB}$$

To facilitate the comparison of PCM with other types of modulation, the SNRs of FM and DSB-SC are superimposed on the SNR of PCM in Fig. 12.14. The theoretical bandwidth expansion ratio for PCM is $B_{\text{PCM}}/B = n$. In practice, this can be achieved by using duobinary signaling. Today's PCM systems use bipolar signaling, however, requiring $B_{\text{PCM}}/B = 2n$. Moreover, P_e in Eq. (12.62) is valid only for polar signaling. Bipolar signaling requires twice as much power. Hence, the plot in Fig. 12.14 is valid for bipolar signaling if 3 dB is added to each value of γ.

From Eq. (12.64b) we have

$$\frac{S_o}{N_o} = 3\left(\frac{\overline{\mathrm{m}^2}}{m_p^2}\right)2^{2B_{\text{PCM}}/kB} \tag{12.65}$$

where $1 \le k \le 2$. For duobinary $k = 1$, and for bipolar $k = 2$.

It is clear from Eq. (12.65) that in PCM, the output SNR increases exponentially with the transmission bandwidth. Compare this with angle modulation, where the SNR increases as a square of the transmission bandwidth. In angle modulation, doubling the transmission bandwidth quadruples the output SNR. From Eq. (12.64c) we see that in PCM, increasing n by 1 quadruples the SNR. But increasing n by 1 increases the bandwidth only by the fraction $1/n$. For $n = 8$, a mere 12.5% increase in the transmission bandwidth quadruples the SNR. Therefore in PCM, the exchange of SNR for bandwidth is much more efficient than in angle modulation. This is particularly evident for large values of n, as can be seen from Fig. 12.14. For smaller values of n, the difference between FM and PCM is not as impressive as for large values of n.*

Actually, to do full justice to PCM, we must consider the use of regenerative repeaters, which cannot be used for angle modulation. The results for PCM derived thus far apply for a single link. If k regenerative repeaters are used along the path, the noise power over each link is reduced by the factor k, and hence P_e, the error probability over each link, is

$$P_e = Q\left(\sqrt{\frac{k\gamma}{n}}\right)$$

But because k links are now in tandem, the overall error probability is kP_e (see Example 10.7), and

$$\frac{S_o}{N_o} = \frac{3(2^{2n})}{1 + 4(2^{2n} - 1)kQ\left(\sqrt{k\gamma/n}\right)}\left(\frac{\overline{m^2}}{m_p^2}\right) \tag{12.66}$$

To maintain a given S_o/N_o, the value of γ decreases as the number of repeaters k, increases. Figure 12.15 shows the γ vs. k needed to maintain $S_o/N_o = 49.5$ dB for $n = 8$.

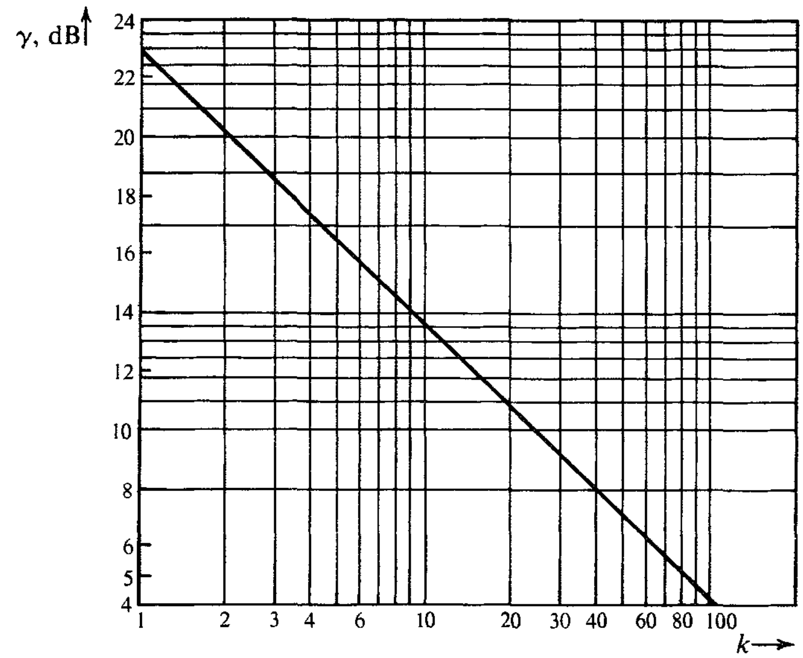


Figure 12.15 Power reduction as a function of the number of repeaters in PCM.

---

* The FM plots in Fig. 12.14 are without preemphasis and deemphasis.

Companded PCM

The output SNR of PCM is proportional to $\overline{m^2}/m_p^2$, where $\overline{m^2}$ is the power of the baseband signal $m(t)$ and m_p is the peak value of $m(t)$. It may appear that $\overline{m^2}/m_p^2$ will remain more or less constant regardless of the speech level, because $\overline{m^2}$ is proportional to m_p^2. Unfortunately, m_p^2 is a constant of the quantizer with a quantization range $(-m_p,\ m_p)$. Once a quantizer is designed, m_p is fixed, and $\overline{m^2}/m_p^2$ is proportional to the speech signal power $\overline{m^2}$ only. This can vary from talker to talker (or even for the same talker) by as much as 40 dB, causing the output SNR to vary widely. This problem can be mitigated, and a relatively constant SNR over a large dynamic range of $\overline{m^2}$ can be obtained, either by nonuniform quantization or by signal companding. Both methods are equivalent, but the latter is simpler to implement. In this method, the signal amplitudes are nonlinearly compressed.

Figure 6.12 shows the input-output characteristics of the two most commonly used compressors (the μ-law and the A-law). For convenience, let us denote

$$x = \frac{m}{m_p}$$

Clearly, the peak value of x is 1 (when $m = m_p$). Moreover, the peak value of the compressor output y is also 1 (occurring when $m = m_p$). Thus, x and y are the normalized input and output of the compressor, each with unit peak value (Fig. 12.16). The input-output characteristics have an odd symmetry about $x = 0$. For convenience, we have only shown the region $x \geq 0$. The output signal samples in the range $(-1,\ 1)$ are uniformly quantized into L levels, with a quantization interval of $2/L$. Figure 12.16 shows the jth quantization interval for the output y as well as the input x. All input sample amplitudes that lie in the range Δ_j are mapped into y_j. For the input sample value x in the range Δ_j, the quantization error is $q = (x - x_j)$, and

$$2\int_{x_j-(\Delta_j/2)}^{x_j+(\Delta_j/2)} (x - x_j)^2 p_x(x)\, dx$$

is the part of $\overline{q^2}$ (the mean square quantizing error) contributed by x in the region Δ_j. The factor 2 appears because there is an equal contribution from negative amplitudes of x centered at $-x_j$. Thus,

$$\overline{q^2} = 2\sum_j \int_{x_j-(\Delta_j/2)}^{x_j+(\Delta_j/2)} (x - x_j)^2 p_x(x) dx$$

Because $L \gg 1$, the quantizing interval $(2/L)$ and Δ_j are very small, and $p_x(x)$ can be assumed to be constant over each interval. Hence,

$$\overline{q^2} = 2\sum_j p_x(x_j) \int_{x_j-(\Delta_j/2)}^{x_j+(\Delta_j/2)} (x - x_j)^2\, dx$$

$$= 2\sum_j \frac{p_x(x_j)\,\Delta_j^3}{12} \tag{12.67}$$

Because $2/L$ and Δ_j are very small, the compression characteristics can be assumed to be linear over each Δ_j, and

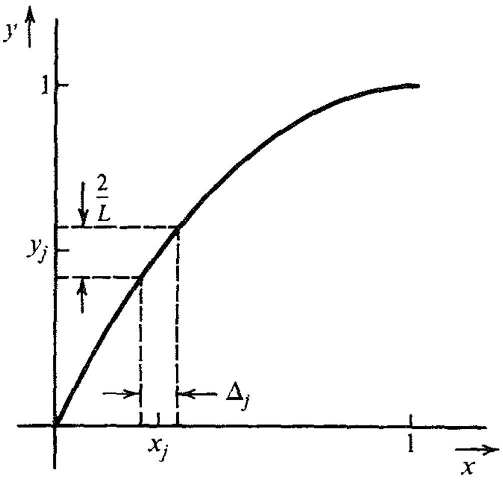


Figure 12.16 Input-output characteristic of a PCM compressor.

$$\dot{y}(x_j) \simeq \frac{2/L}{\Delta_j}$$

Substituting this in Eq. (12.67), we have

$$\overline{\mathrm{q}^2} \simeq \frac{2}{3L^2} \sum_j \frac{p_{\mathrm{x}}(x_j)}{[\dot{y}(x_j)]^2} \Delta_j$$

For L large enough, the preceding sum can be approximated by an integral

$$\overline{\mathrm{q}^2} \simeq \frac{2}{3L^2} \int_0^1 \frac{p_{\mathrm{x}}(x)}{[\dot{y}(x)]^2} dx \tag{12.68}$$

For the μ-law [Eq. (6.17a)],

$$y = \frac{\ln(1+\mu x)}{\ln(1+\mu)} \qquad 0 \le x \le 1$$

and

$$\dot{y}(x) = \frac{\mu}{\ln(1+\mu)} \left(\frac{1}{1+\mu x}\right)$$

and

$$\overline{\mathrm{q}^2} = \left(\frac{2}{3L^2}\right) \left[\frac{\ln(1+\mu)}{\mu}\right]^2 \int_0^1 (1+\mu x)^2 p_{\mathrm{x}}(x)\, dx \tag{12.69}$$

If $p_{\mathrm{x}}(x)$ is symmetrical about $x = 0$,

$$\sigma_{\mathrm{x}}^2 = 2 \int_0^1 x^2 p_{\mathrm{x}}(x)\, dx \tag{12.70a}$$

and $\overline{|\mathrm{x}|}$, the mean of the rectified x, is

$$\overline{|\mathrm{x}|} = 2 \int_0^1 x p_{\mathrm{x}}(x)\, dx \tag{12.70b}$$

We can express $\overline{\mathrm{q}^2}$ as

$$\overline{\mathrm{q}^2} = \left[\frac{\ln(1+\mu)}{\mu}\right]^2 \left[\frac{1+\mu^2\sigma_{\mathrm{x}}^2 + 2\mu\overline{|\mathrm{x}|}}{3L^2}\right] \tag{12.71a}$$

$$= \frac{[\ln(1+\mu)]^2}{3L^2} \cdot \left(\sigma_{\mathrm{x}}^2 + \frac{2\overline{|\mathrm{x}|}}{\mu} + \frac{1}{\mu^2}\right) \tag{12.71b}$$

Recall that $\overline{\mathrm{q}^2}$ in Eqs. (12.71) is the normalized quantization error. The actual error is $m_p^2\,\overline{\mathrm{q}^2}$. The normalized output signal is $\mathrm{x}(t)$, and, hence, the normalized output power $S_o = \sigma_{\mathrm{x}}^2 = \overline{\mathrm{m}^2}/m_p^2$. The actual S_o will be $m_p^2\,\sigma_{\mathrm{x}}^2$. Hence,

$$\frac{S_o}{N_o} = \frac{\sigma_{\mathrm{x}}^2}{\overline{\mathrm{q}^2}} = \frac{3L^2}{[\ln(1+\mu)]^2} \frac{\sigma_{\mathrm{x}}^2}{\left(\sigma_{\mathrm{x}}^2 + 2\overline{|\mathrm{x}|}/\mu + 1/\mu^2\right)} \tag{12.72a}$$

$$= \frac{3L^2}{[\ln(1+\mu)]^2} \frac{1}{\left(1 + 2\overline{|\mathrm{x}|}/\mu\sigma_{\mathrm{x}}^2 + 1/\mu^2\sigma_{\mathrm{x}}^2\right)} \tag{12.72b}$$

To get an idea of the relative importance of the various terms in the parentheses in Eq. (12.72b), we note that x is an RV distributed in the range $(-1,\ 1)$. Hence, σ_{x}^2 and $\overline{|\mathrm{x}|}$ are both less than 1, and $\overline{|\mathrm{x}|}/\sigma_{\mathrm{x}}$ is typically in the range of 0.7 to 0.9. The values of μ used in practice are greater than 100. For example, the D2 channel bank used in conjunction with the T1 carrier system has $\mu = 255$. Thus, the second and third terms in the parentheses in Eq. (12.72b) are small compared to 1 if σ_{x}^2 is not too small, and so

$$\frac{S_o}{N_o} \simeq \frac{3L^2}{[\ln(1+\mu)]^2} \tag{12.72c}$$

which is independent of σ_{x}^2. The exact expression in Eq. (12.72b) has a weak dependence on σ_{x}^2 over a broad range of σ_{x}. Note that the SNR in Eq. (12.72b) also depends on the signal statistics $\overline{|\mathrm{x}|}$ and σ_{x}^2. But for most of the practical PDFs, $\overline{|\mathrm{x}|}/\sigma_{\mathrm{x}}$ is practically the same (in the range of 0.7 to 0.9). Hence, S_o/N_o depends only on σ_{x}^2. This means the plot of S_o/N_o vs. σ_{x}^2 will be practically independent of the PDF of x. Figure 12.17 shows the plot of S_o/N_o vs. σ_{x}^2 for two different PDFs; Laplacian and gaussian (see the next example). It can be seen that there is hardly any difference between the two curves.

Because $\mathrm{x} = \mathrm{m}/m_p$, $\sigma_{\mathrm{x}}^2 = \sigma_{\mathrm{m}}^2/m_p^2$, and $\overline{|\mathrm{x}|} = \overline{|\mathrm{m}|}/m_p$, Eq. (12.72a) becomes

$$\frac{S_o}{N_o} = \frac{3L^2}{[\ln(1+\mu)]^2} \left[\frac{\sigma_{\mathrm{m}}^2/m_p^2}{\sigma_{\mathrm{m}}^2/m_p^2 + 2\overline{|\mathrm{m}|}/\mu m_p + 1/\mu^2}\right] \tag{12.73}$$

One should be careful in interpreting m_p in Eq. (12.73). Once the system is designed for some $\mathrm{m}(t)$, m_p is fixed. Hence, m_p is a constant of the system, not of the signal $\mathrm{m}(t)$ that may be subsequently transmitted.

EXAMPLE 12.9 A voice signal amplitude PDF can be closely modeled by the Laplace density[*]

---

[*] A better but more complex model for speech signal amplitude m is the gamma density[6]

$$p_{\mathrm{m}}(m) = \sqrt{\frac{k}{4\pi|m|}}\, e^{-k|m|}$$

$$p_{\mathrm{m}}(m) = \frac{1}{\sigma_{\mathrm{m}}\sqrt{2}} e^{-\sqrt{2}|m|/\sigma_{\mathrm{m}}}$$

For a voice PCM system with $n = 8$ and $\mu = 255$, find and sketch the output SNR as a function of the normalized voice power $\sigma_{\mathrm{m}}^2/m_p^2$.

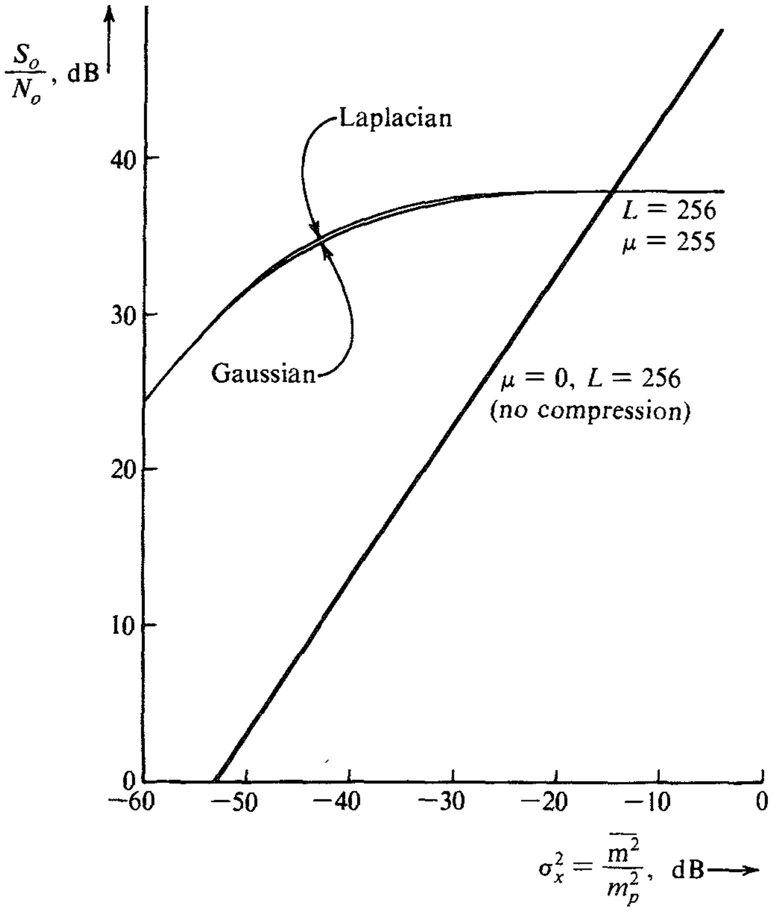


Figure 12.17 PCM performance with and without companding.

It is straightforward to show that the variance of this Laplace PDF is σ_{m}^2. In practice, the speech amplitude will be limited by either 3σ or 4σ loading. In either case, the probability of observing m beyond this limit will be negligible, and in computing $\overline{|\mathrm{m}|}$ (etc.), we may use the limits 0 to ∞,

$$\overline{|\mathrm{m}|} = 2\int_0^{\infty} \frac{m}{\sigma_{\mathrm{m}}\sqrt{2}} e^{-\sqrt{2}m/\sigma_x}\, dm = 0.707\sigma_{\mathrm{m}}$$

Hence, from Eq. (12.73),

$$\frac{S_o}{N_o} = \frac{6394(\sigma_{\mathrm{m}}^2/m_p^2)}{(\sigma_{\mathrm{m}}^2/m_p^2) + 0.00555(\sigma_{\mathrm{m}}/m_p) + 1.53 \times 10^{-5}} \tag{12.74}$$

This is plotted as a function of $(\sigma_{\mathrm{m}}^2/m_p^2)$ in Fig. 12.17.

EXAMPLE 12.10 Repeat Example 12.9 for the gaussian m(t).

In this case,

$$\overline{|\mathrm{m}|} = 2\int_0^{\infty} \frac{m}{\sigma_{\mathrm{m}}\sqrt{2\pi}} e^{-m^2/2\sigma_{\mathrm{m}}^2}\, dm = 0.798\sigma_{\mathrm{m}}$$

and

$$\frac{S_o}{N_o} = \frac{6394(\sigma_{\mathrm{m}}^2/m_p^2)}{(\sigma_{\mathrm{m}}^2/m_p^2) + 0.0063(\sigma_{\mathrm{m}}/m_p) + 1.53 \times 10^{-5}} \tag{12.75}$$

The SNR here is nearly the same as that in Eq. (12.74). The plot of SNR vs. $\sigma_{\mathrm{m}}^2/m_p^2$ (Fig. 12.17) is practically indistinguishable from that in Example 12.9.

12.5 OPTIMUM PREEMPHASIS-DEEMPHASIS SYSTEMS

It is possible to increase the output SNR by deliberate distortion of the transmitted signal (preemphasis) and the corresponding compensation (deemphasis) at the receiver. For intuitive understanding of this process, consider a case of white channel noise and a signal m(t) whose PSD decreases with frequency. In this case, we can boost the high-frequency components of m(t) at the transmitter (preemphasis). Because the signal has relatively less power at high frequencies, this preemphasis will require only a small increase in transmitted power.* At the receiver, the high-frequency components are attenuated (or deemphasized) in order to undo the preemphasis at the transmitter. This will restore the useful signal to its original form. It is an entirely different story with the channel noise. Because the noise is added after the transmitter, it does not undergo preemphasis. At the receiver, however, it does undergo deemphasis (i.e., attenuation of high-frequency components). Thus, at the receiver output, the signal power is restored but the noise power is reduced. The output SNR is therefore increased.

We shall first consider a baseband system and then extend the discussion to modulated systems. A baseband system with a preemphasis filter $H_p(\omega)$ at the transmitter and the corresponding complementary deemphasis filter $H_d(\omega)$ at the receiver is shown in Fig. 12.18. The channel transfer function is $H_c(\omega)$, and the PSD of the input signal m(t) is $S_{\mathrm{m}}(\omega)$. We shall determine the optimum preemphasis-deemphasis (PDE) filters $H_p(\omega)$ and $H_d(\omega)$ required for distortionless transmission of the signal m(t).

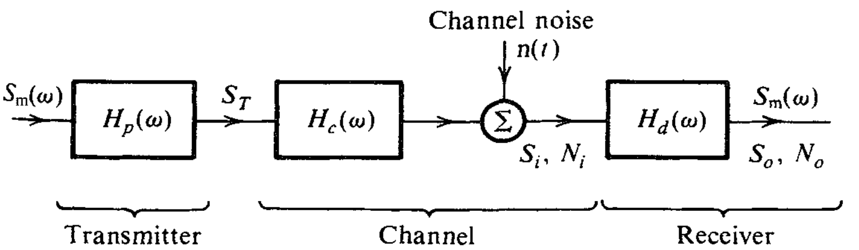


Figure 12.18 Optimum PDE in a baseband system.

---

* Actually, the transmitted power is maintained constant by attenuating the preemphasized signal slightly.

For distortionless transmission,

$$|H_p(\omega)H_c(\omega)H_d(\omega)| = G \quad \text{(a constant)} \tag{12.76a}$$

and

$$\theta_p(\omega) + \theta_c(\omega) + \theta_d(\omega) = -\omega t_d \tag{12.76b}$$

We want to maximize the output SNR, S_o/N_o, for a given transmitted power S_T.

Referring to Fig. 12.18, we have

$$S_T = \int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)|H_p(\omega)|^2\, df \tag{12.77a}$$

Because $|H_p(\omega)H_c(\omega)H_d(\omega)| = G$, the signal power S_o at the receiver output is

$$S_o = G^2 \int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)\, df \tag{12.77b}$$

The noise power N_o at the receiver output is

$$N_o = \int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H_d(\omega)|^2\, df \tag{12.77c}$$

Thus,

$$\frac{S_o}{N_o} = \frac{G^2 \int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)\, df}{\int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H_d(\omega)|^2\, df} \tag{12.78}$$

We wish to maximize this ratio subject to the condition in Eq. (12.77a) with S_T as a given constant. We can include this constraint by multiplying the numerator and the denominator of the right-hand side of Eq. (12.78) by the left-hand side and the right-hand side, respectively, of Eq. (12.77a). This gives

$$\frac{S_o}{N_o} = \frac{G^2 S_T \int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)\, df}{\int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H_d(\omega)|^2\, df \int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)|H_p(\omega)|^2\, df} \tag{12.79}$$

The numerator of the right-hand side of Eq. (12.79) is fixed. Hence, to maximize S_o/N_o, we need only minimize the denominator of the right-hand side of Eq. (12.79). To do this, we use the Schwarz inequality (Appendix B),

$$\begin{aligned} &\int_{-\infty}^{\infty} S_{\mathrm{m}}(\omega)|H_p(\omega)|^2\, df \int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H_d(\omega)|^2\, df \\ &\qquad \geq \left| \int_{-\infty}^{\infty} [S_{\mathrm{m}}(\omega)S_{\mathrm{n}}(\omega)]^{1/2} |H_p(\omega)H_d(\omega)|\, df \right|^2 \end{aligned} \tag{12.80}$$

The equality holds only if

$$S_{\mathrm{m}}(\omega)|H_p(\omega)|^2 = K^2 S_{\mathrm{n}}(\omega)|H_d(\omega)|^2 \tag{12.81}$$

where K is an arbitrary constant. Thus to maximize S_o/N_o, Eq. (12.81) must be satisfied. Substitution of Eq. (12.76a) into Eq. (12.81) yields

$$|H_p(\omega)|^2_{\mathrm{opt}} = GK \frac{\sqrt{S_{\mathrm{n}}(\omega)/S_{\mathrm{m}}(\omega)}}{|H_c(\omega)|} \tag{12.82a}$$

$$|H_d(\omega)|^2_{\text{opt}} = \frac{G}{K}\,\frac{\sqrt{S_{\text{m}}(\omega)/S_{\text{n}}(\omega)}}{|H_c(\omega)|} \tag{12.82b}$$

The constant K is found by substituting Eq. (12.82a) into Eq. (12.77a) as

$$K = \frac{S_T}{G\int_{-\infty}^{\infty}[\sqrt{S_{\text{m}}(\omega)S_{\text{n}}(\omega)}/|H_c(\omega)|]\,df} \tag{12.82c}$$

Substitution of this value of K into Eqs. (12.82a, b) yields

$$|H_p(\omega)|^2_{\text{opt}} = \frac{S_T\sqrt{S_{\text{n}}(\omega)/S_{\text{m}}(\omega)}}{|H_c(\omega)|\int_{-\infty}^{\infty}[\sqrt{S_{\text{m}}(\omega)S_{\text{n}}(\omega)}/|H_c(\omega)|]\,df} \tag{12.83a}$$

$$|H_d(\omega)|^2_{\text{opt}} = \frac{G^2\int_{-\infty}^{\infty}[\sqrt{S_{\text{m}}(\omega)S_{\text{n}}(\omega)}/|H_c(\omega)|]\,df}{S_T|H_c(\omega)|\sqrt{S_{\text{n}}(\omega)/S_{\text{m}}(\omega)}} \tag{12.83b}$$

The output SNR under optimum conditions is given by Eq. (12.79) with its denominator replaced with the right-hand side of Eq. (12.80). Now substituting $|H_p(\omega)H_d(\omega)| = G/|H_c(\omega)|$ yields

$$\left(\frac{S_o}{N_o}\right)_{\text{opt}} = \frac{S_T\int_{-\infty}^{\infty}S_{\text{m}}(\omega)\,df}{\left(\int_{-\infty}^{\infty}[\sqrt{S_{\text{m}}(\omega)S_{\text{n}}(\omega)}/|H_c(\omega)|]\,df\right)^2} \tag{12.83c}$$

Equations (12.83a) and (12.83b) give the magnitudes of the optimum filters $H_p(\omega)$ and $H_d(\omega)$. The phase functions must be chosen to satisfy the condition of distortionless transmission [Eq. (12.76b)].

Observe that the preemphasis filter in Eq. (12.82a) boosts frequency components where the signal is weak and suppresses frequency components where the signal is strong. The deemphasis filter in Eq. (12.82b) does exactly the opposite. Thus, the signal is unchanged but the noise is reduced.

EXAMPLE 12.11 Consider the case with $\alpha = 1400\pi$,

$$S_{\text{m}}(\omega) = \begin{cases} \dfrac{C}{\omega^2+\alpha^2} & |\omega| \le 8000\pi \\ 0 & |\omega| \ge 8000\pi \end{cases} \tag{12.84a}$$

The channel noise is white with PSD

$$S_{\text{n}}(\omega) = \frac{\mathcal{N}}{2} \tag{12.84b}$$

The channel is assumed to be ideal [$H_c(\omega) = 1$ and $G = 1$] over the band of interest (0 to 4000 Hz).

Without preemphasis-deemphasis, we have

$$\begin{aligned} S_o &= \frac{1}{2\pi}\int_{-8000\pi}^{8000\pi} S_{\text{m}}(\omega)\,d\omega \\ &= \frac{1}{\pi}\int_0^{8000\pi}\frac{C}{\omega^2+\alpha^2}\,d\omega \qquad \alpha = 1400\pi \\ &= 10^{-4}C \end{aligned}$$

Also, because $G = 1$, the transmitted power $S_T = S_o$,

$$S_o = S_T = 10^{-4}C$$

and the noise power without preemphasis-deemphasis is

$$N_o = \mathcal{N}B = 4000\mathcal{N}$$

Therefore,

$$\frac{S_o}{N_o} = 2.5 \times 10^{-8}\frac{C}{\mathcal{N}} \tag{12.85}$$

The optimum transmitting and receiving filters are given by [Eqs. (12.83a and b)]

$$|H_p(\omega)|^2 = \frac{10^{-4}\sqrt{\omega^2+\alpha^2}}{\int_{-\infty}^{\infty}\left(1/\sqrt{\omega^2+\alpha^2}\right)df} = \frac{1.286\sqrt{\omega^2+\alpha^2}}{10^4} \qquad |\omega| \leq 8000\pi \tag{12.86a}$$

$$|H_d(\omega)|^2 = \frac{10^4\int_{-\infty}^{\infty}\left(1/\sqrt{\omega^2+\alpha^2}\right)df}{\sqrt{\omega^2+\alpha^2}} = \frac{0.778 \times 10^4}{\sqrt{\omega^2+\alpha^2}} \qquad |\omega| \leq 8000\pi \tag{12.86b}$$

The output SNR using optimum preemphasis and deemphasis in found from Eq. (12.83c) as

$$\begin{aligned}\left(\frac{S_o}{N_o}\right)_{\text{opt}} &= \frac{(10^{-4}C)^2}{(\mathcal{N}C/2)\left[\int_{-4000}^{4000}\left[1/\sqrt{4\pi^2f^2+(1400\pi)^2}\right]df\right]^2}\\ &= 3.3 \times 10^{-8}\frac{C}{\mathcal{N}}\end{aligned} \tag{12.87}$$

Comparison of Eq. (12.85) with Eq. (12.87) shows that preemphasis-deemphasis has increased the output SNR by a factor of 1.32.

Optimum Preemphasis-Deemphasis in AM Systems

Because the channel noise in a DSB system is in the band $\omega_c \pm 2\pi B$, the PDE should be carried out in this band. This means PDE filters $H_p(\omega)$ and $H_d(\omega)$ are bandpass filters located as shown in Fig. 12.19. The optimization is localized to the subsystem shown in the dashed box. This subsystem is identical to that in Fig. 12.18. Hence, $H_p(\omega)$ and $H_d(\omega)$ can be obtained from Eqs. (12.83a, b). However, because the signal PSD at the input of the subsystem in Fig. 12.19 is $\frac{1}{2}[S_m(\omega+\omega_c) + S_m(\omega-\omega_c)]$, we should use this PSD in place of $S_m(\omega)$ in Eqs. (12.83a, b). The same argument applies to SSB and VSB systems. In these cases, $H_p(\omega)$ and $H_d(\omega)$ can be computed from Eqs. (12.83a, b) by replacing $S_m(\omega)$ with the appropriate SSB or VSB PSD.

Optimum Preemphasis-Deemphasis in Angle Modulation

We shall consider the case of PDE in the baseband, that is, preemphasis before modulation and deemphasis after demodulation (Fig. 12.20a). The channel is assumed to be ideal, that is, $H_c(\omega) = 1$.

The problem here is very different from the baseband communication case. In the baseband case, the preemphasis filter changes the transmitted power but not its bandwidth,

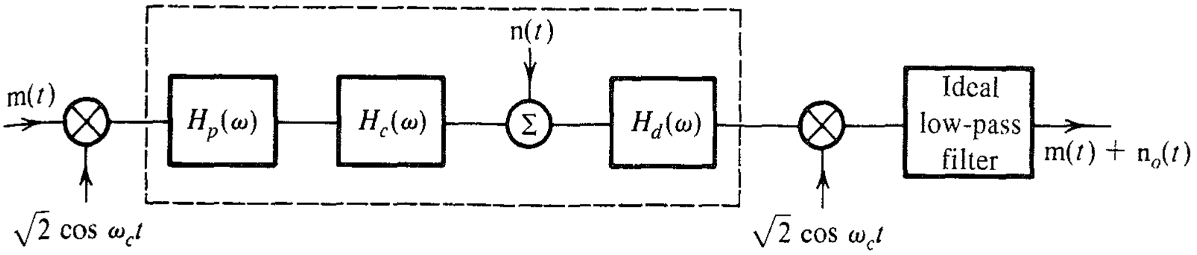
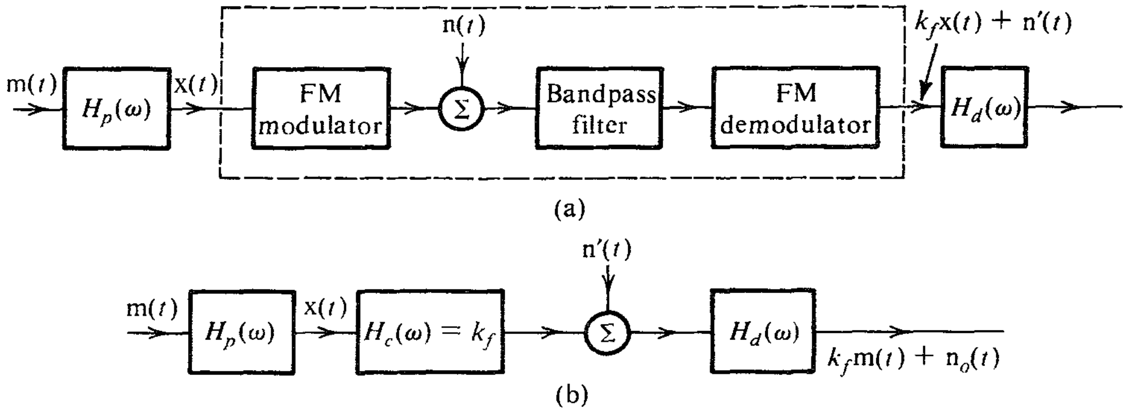

m(t)
$\sqrt{2}\cos\omega_c t$
$H_p(\omega)$
$H_c(\omega)$
n(t)
Σ
$H_d(\omega)$
$\sqrt{2}\cos\omega_c t$
Ideal low-pass filter
m(t) + $n_o(t)$

Figure 12.19 Optimum PDE in an amplitude-modulated system.

and optimization was performed under the constraint of a fixed transmitted power (because bandwidth remained fixed, it was of no concern). In FM, on the other hand, the transmitted power is always fixed ($A^2/2$), but the transmission bandwidth is modified by the preemphasis filter. But optimization must be performed under the constraint of a fixed transmission bandwidth. Which bandwidth should be used? The conventional ($2\Delta f$) or the mean square? A look at Eq. (12.36) shows that the SNR in terms of conventional bandwidth involves the constant m_p. When m(t) is passed through $H_p(\omega)$, m_p will change, and no simple relationship exists between m(t), $H_p(\omega)$, and the new m_p. The problem is mathematically intractable. No such difficulty appears when we use Eq. (12.47), which uses the mean square bandwidth. Hence, we shall optimize the ratio S_o/N_o with the constraint of a fixed mean square bandwidth of transmission. Fortunately, this constraint turns out to be similar to that used in the baseband system. This can be seen from Eq. (12.44):

$$\overline{B_{\text{FM}}^2} = \frac{1}{4\pi^2} k_f^2 \, \overline{\text{m}^2}$$

Thus to maintain $\overline{B_{\text{FM}}^2}$ fixed, $\overline{\text{m}^2}$, the power of the modulating signal, must be the same with or without preemphasis. This is exactly the constraint of the baseband system [Eq. (12.77a)]. If the input to the system in the dotted box (Fig. 12.20a) is x(t), its output is k_fx(t) plus the parabolic noise n′(t) with the PSD in Eq. (12.33). Hence, the system in Fig. 12.20b is the equivalent of the system in Fig. 12.20a. Our problem is the optimization of the output SNR with the constraint that the output of $H_p(\omega)$ has a given mean square value. This problem is


m(t)
$H_p(\omega)$
x(t)
FM modulator
n(t)
Σ
Bandpass filter
FM demodulator
k_fx(t) + n′(t)
$H_d(\omega)$
m(t)
$H_p(\omega)$
x(t)
$H_c(\omega) = k_f$
n′(t)
Σ
$H_d(\omega)$
k_fm(t) + $n_o(t)$

(a)

(b)

Figure 12.20 (a) Optimum PDE in FM system. (b) Equivalent of system in part (a).

identical to the optimization of the baseband system with $G = H_c(\omega) = k_f$. Hence, from Eqs. (12.83) with $S_n(\omega) = \mathcal{N}\omega^2/A^2$ and $S_T = \overline{m^2}$, we have

$$|H_p(\omega)|^2_{\text{opt}} = \frac{\overline{m^2}\omega}{\sqrt{S_m(\omega)}\int_{-\infty}^{\infty}\sqrt{\omega^2 S_m(\omega)}\,df} \tag{12.88a}$$

$$|H_d(\omega)|^2_{\text{opt}} = \frac{\sqrt{S_m(\omega)}\int_{-\infty}^{\infty}\sqrt{\omega^2 S_m(\omega)}\,df}{\overline{m^2}\omega} \tag{12.88b}$$

$$\left(\frac{S_o}{N_o}\right)_{\text{opt}} = \frac{\left(\frac{A^2 k_f^2}{4\pi^2\mathcal{N}}\right)\overline{m^2}\int_0^B S_m(\omega)\,df}{2\left[\int_0^B f\sqrt{S_m(\omega)}\,df\right]^2} \tag{12.88c}$$

Because $\gamma = A^2/2\mathcal{N}B$,

$$\left(\frac{S_o}{N_o}\right)_{\text{opt}} = \left[\frac{B^3\,\overline{m^2}}{6\left[\int_0^B f\sqrt{S_m(\omega)}\,df\right]^2}\right]\left(\frac{3B_{\text{FM}}^2}{B^2}\gamma\right) \tag{12.88d}$$

Comparison of Eq. (12.88d) with Eq. (12.47) shows that PDE in FM improves the SNR by the factor inside the brackets on the right-hand side of Eq. (12.88d). Optimum preemphasis and deemphasis is not used in commercial FM broadcasting for historical and practical reasons. The relatively simple suboptimum scheme, discussed in Sec. 5.5, is used instead.

REFERENCES

1. P. F. Panter, *Modulation, Noise, and Spectral Analysis*, McGraw-Hill, New York, 1965.
2. N. Abramson, "Bandwidth and Spectra of Phase- and Frequency-Modulated Waves," *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 407–414, Dec. 1963.
3. S. O. Rice, "Mathematical Analysis of Random Noise," *Bell Syst. Tech. J.*, vol. 23, pp. 282–332, July 1944; vol. 24, pp. 46–156, Jan. 1945.
4. A. J. Viterbi, *Principles of Coherent Communication*, McGraw-Hill, New York, 1966.
5. H. E. Rowe, *Signals and Noise in Communication Systems*, Van Nostrand, Princeton, NJ, 1965.
6. M. D. Paez and T. H. Glissom, "Minimum Mean Square Error Quantization in Speech, PCM, and DPCM Systems," *IEEE Trans. Commun. Technol.*, vol. COM-20, pp. 225–230, April 1972.

PROBLEMS

12.1-1 A certain telephone channel has $H_c(\omega) \simeq 10^{-3}$ over the signal band. The message signal PSD is $S_m(\omega) = \beta\,\text{rect}\,(\omega/2\alpha)$, with $\alpha = 8000\pi$. The channel noise PSD is $S_n(\omega) = 10^{-8}$. If the output SNR at the receiver is required to be at least 30 dB, what is the minimum transmitted power required? Calculate the value of β corresponding to this power.

12.1-2 A signal $m(t)$ with PSD $S_m(\omega) = \beta\,\text{rect}\,(\omega/2\alpha)$ and $\alpha = 8000\pi$ is transmitted over a telephone channel with transfer function $H_c(\omega) = 10^{-3}/(j\omega + \alpha)$. The channel noise PSD is $S_n(\omega) = 10^{-10}$. To compensate for the channel distortion, the receiver filter transfer function is chosen to be

$$H_d(\omega) = \left(\frac{j\omega + \alpha}{\alpha}\right) \text{rect}\left(\frac{\omega}{2\alpha}\right)$$

The receiver output SNR is required to be at least 35 dB. Determine the minimum required value of β and the corresponding transmitted power S_T and the power S_i received at the receiver input.

12.2-1 For a DSB-SC system with a channel noise PSD of $S_n(\omega) = 10^{-10}$ and a baseband signal of bandwidth 4 kHz, the receiver output SNR is required to be at least 30 dB. The receiver is as shown in Fig. 12.3.

(a) What must be the signal power S_i received at the receiver input?

(b) What is the receiver output noise power N_o?

(c) What is the minimum transmitted power S_T if the channel transfer function is $H_c(\omega) = 10^{-4}$ over the transmission band?

12.2-2 Repeat Prob. 12.2-1 for SSB-SC.

12.2-3 Determine the output SNR of each of the two quadrature multiplexed channels and compare the results with those of DSB-SC and SSB-SC.

12.2-4 Assume $[m(t)]_{\max} = -[m(t)]_{\min} = m_p$.

(a) Show that for AM,

$$m_p = \mu A$$

(b) Show that the output SNR for AM [Eq. (12.14)] can be expressed as

$$\frac{S_o}{N_o} = \frac{\mu^2}{k^2 + \mu^2}\gamma$$

where $k^2 = m_p^2/\overline{m^2}$.

(c) Using the result in part **(b)**, show that for tone modulation with $\mu = 1$, $S_o/N_o = \gamma/3$.

(d) Show that if S_T and S'_T are the AM and DSB-SC transmitted powers, respectively, required to attain a given output SNR, then

$$S_T \simeq k^2 S'_T \qquad \text{for } \mu = 1 \quad \text{and} \quad k^2 \gg 1.$$

12.2-5 A gaussian baseband random process $m(t)$ is transmitted by AM.

(a) Show that for 3σ loading (that is, $m_p = 3\sigma$), the output SNR is $\gamma/10$ when $\mu = 1$.

(b) Show that for 3σ loading and $\mu = 0.5$, the output SNR $\simeq \gamma/36$.

12.2-6 In many communication systems, the transmitted signal is limited by peak power rather than by average power. Under such a limitation, AM fares much worse than DSB-SC or SSB-SC. Show that for tone modulation for a fixed peak power transmitted, the output SNR of AM is 6 dB below that of DSB-SC and 9 dB below that of SSB-SC.

12.2-7 Determine γ_{thresh} in AM with $\mu = 1$ if the onset of the threshold is when $E_n > A$ with probability 0.01, where E_n is the noise envelope. Assume the modulating signal $m(t)$ to be gaussian and use 4σ loading.

12.3-1 For an FM communication system with $\beta = 2$ and white channel noise with PSD $S_n(\omega) = 10^{-10}$, the output SNR is found to be 28 dB. The baseband signal $m(t)$ is gaussian and band-limited to 15 kHz, and 3σ loading is used. The demodulator constant $\alpha = 10^{-4}$. This means that the FM demodulator output is $\alpha\dot{\psi}(t)$ when the input is $A\cos[\omega_c t + \psi(t)]$. In the present case, the signal at the demodulator output is $\alpha k_f m(t)$. The output noise is also multiplied by α.

(a) Determine the received signal power S_i.

(b) Determine the output signal power S_o.

(c) Determine the output noise power N_o.

12.3-2 For the modulating signal $m(t)$ shown in Fig. P12.3-2, show that PM is superior to FM by a factor $3\pi^2/4$ from the SNR point of view. *Hint:* Assume the bandwidth of $m(t)$ to be the frequency of its third harmonic.

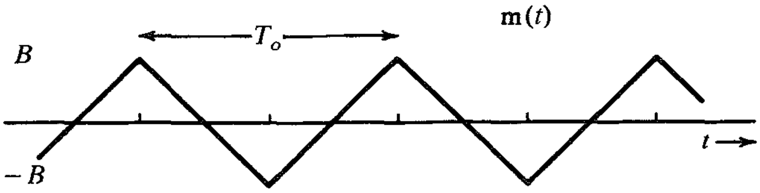


Figure P12.3-2

12.3-3 For a modulating signal $m(t) = \cos^3 \omega_o t$, show that PM is superior to FM by a factor of 2.25 from the SNR point of view. *Hint:* Determine m'_p by setting $\ddot{m}(t) = 0$.

12.3-4 For $m(t) = a_1 \cos\omega_1 t + a_2 \cos\omega_2 t$, show that PM is superior to FM from the SNR point of view when $(1 + xy)^2 < (1 + x)^2/3$, where $x = a_1/a_2$ and $y = \omega_1/\omega_2$. *Hint:* Use Eq. (12.39).

12.3-5 Show that $\overline{B_m^2}$ in Eq. (12.42a) can also be expressed directly in the time domain as

$$\overline{B_m^2} = \frac{\int_{-\infty}^{\infty} [\dot{m}(t)]^2\,dt}{\int_{-\infty}^{\infty} m^2(t)\,dt}$$

12.3-6 Show that when the modulating signal $m(t)$ has a Butterworth PSD, that is,

$$S_m(\omega) = \frac{1}{1 + (\omega/2\pi f_o)^{2k}}$$

then the mean square bandwidth $\overline{B_m^2}$ is given by

$$\overline{B_m^2} = f_o^2 \frac{\sin(\pi/2k)}{\sin(3\pi/2k)} \qquad k \geq 2$$

and as $k \to \infty$, $\overline{B_m^2} \to \frac{1}{3}f_o^2$.

12.3-7 A modulating signal PSD is given by

$$S_m(\omega) = \frac{|\omega|}{\sigma^2} e^{-\omega^2/2\sigma^2}$$

The effective bandwidth B of $m(t)$ is that bandwidth which contains χ percent of the total power of $m(t)$. Determine which of the angle modulations is superior if: **(a)** $\chi = 99$; **(b)** $\chi = 95$; **(c)** $\chi = 90$.

12.3-8 In Prob. 12.3-4, FM and PM were compared for a given transmission bandwidth. If we compare them for a given mean square transmission bandwidth, we get a slightly different result. Show that PM is superior to FM if $(1 + x^2y^2) < (1 + x^2)/3$ when they are compared for a given mean square transmission bandwidth. This example shows that two signal spectra with the same mean square bandwidth can have different conventional bandwidths. *Hint*: In this case $S_{\mathrm{m}}(\omega)$ is discrete. Normalize it, and then take the second moment about the origin. Show that

$$\overline{B_{\mathrm{m}}^2} = \frac{a_1^2 f_1^2 + a_2^2 f_2^2}{a_1^2 + a_2^2}$$

12.3-9 In a certain FM system used in space communication, the output SNR is found to be 23.4 dB with $\beta = 2$. The modulating signal $\mathrm{m}(t)$ is gaussian with a bandwidth of 10 kHz, and 3σ loading is used. The system with $\beta = 2$ is in the nonthreshold region of operation. The output SNR is required to be at least 40 dB. Because power is at premium in space communication, it is decided to increase the output SNR by increasing β (i.e., increasing the transmission bandwidth) as much as is possible.

(a) What are the maximum value of β and the corresponding transmission bandwidth that can be used without running into the threshold? What is the corresponding output SNR? *Hint*: Use Eqs. (12.51) and (12.37) to determine the minimum usable γ.

(b) What must be the minimum increase in the transmitted power required to attain an output SNR of 40 dB? What are the corresponding value of β and the transmission bandwidth?

12.3-10 Show that

(a) For tone modulation, the dividing line between narrow-band and wide-band modulation is $\beta = 0.47$.

(b) For a gaussian modulating signal with 3σ loading, the dividing line is at $\beta = 0.55$.

(c) For 4σ loading it is at $\beta = 0.56$.

12.3-11 For FM stereophonic broadcasting (Fig. 5.19), show that the $(L - R)$ channel is about 22 dB noisier than the $L + R$ channel. *Hint:* $L + R$ is the baseband signal. It is preemphasized to obtain $(L+R)'$. The signal $(L-R)$ is preemphasized to obtain $(L-R)'$, which is used to obtain $(L - R)' \cos \omega_c t$. The sum $(L + R)' + (L - R)' \cos \omega_c t$ now frequency-modulates a carrier. At the receiver, after frequency demodulation, $(L + R)'$ and $(L - R)' \cos \omega_c t$ are separated. $(L-R)'$ is deemphasized. $(L+R)' \cos \omega_c t$ is multiplied by $2 \cos \omega_c t$ to obtain $(L-R)'$, which is then deemphasized.

12.4-1 In M-ary PCM, pulses can take M distinct amplitudes (in contrast to two for binary PCM). Show that the signal-to-quantization-noise ratio for M-ary PCM is

$$\frac{S_o}{N_o} = 3M^{2n} \left(\frac{\overline{\mathrm{m}^2}}{m_p^2} \right)$$

12.4-2 A TV signal band-limited to 4.5 MHz is to be transmitted by binary PCM. The receiver output signal-to-quantization-noise ratio is required to be at least 55 dB.

(a) If all brightness levels are assumed to be equally likely, that is, amplitudes of $\mathrm{m}(t)$ are uniformly distributed in the range $(-m_p, m_p)$, find the minimum number of quantization levels L required. Select the nearest value of L to satisfy $L = 2^n$.

(b) For this value of L, compute the receiver output SNR and the transmission bandwidth, assuming the nonthreshold region of operation.

(c) If the output SNR is required to be increased by 6 dB (four times), what are the new value of L and the corresponding transmission bandwidth?

12.4-3 A modulating signal m(t) band-limited to 4 kHz is sampled at a rate of 8000 samples per second. The samples are quantized into 256 levels, binary coded, and transmitted over a channel with $S_n(\omega) = 6.25 \times 10^{-7}$. Each received pulse has energy $E_p = 2 \times 10^{-5}$. Given that $m_p = 1$ and $\overline{m^2} = 1/9$,

(a) Find the output SNR assuming polar signal, and the error probability given in Eq. (12.62).

(b) If the transmitted power is reduced by 10 dB, find the new SNR.

(c) At the reduced power level in part **(b)**, is it possible to increase the output SNR by changing the value of L? Determine the maximum output SNR achievable and the corresponding value of L.

12.4-4 In a PCM channel using k identical regenerative links, we have shown that the error probability P_E of the overall channel is kP_e, where P_e is the error probability of an individual link (see Example 10.7). This shows that P_e is cumulative.

(a) Show that if $k - 1$ links are identical with error probability P_e and the remaining one link has an error probability P_e', then

$$P_E = (k-1)P_e + P_e'$$

(b) For a certain chain of repeaters with $k = 100$ (100 repeaters), it is found that γ over each of the 99 links is 25 dB, and over the remaining link γ is 23 dB. Calculate P_e and P_e' using Eq. (12.62) (with $n = 8$). Now compute P_E and show that P_E is primarily determined by the single weakest link in the chain.

12.4-5 For companded PCM with $n = 8$, $\mu = 255$, and amplitude m uniformly distributed in the range $(-A,\ A)$ $(A \leq m_p)$, show that

$$\frac{S_o}{N_o} = \frac{6394(\sigma_{\mathrm{m}}^2/m_p^2)}{(\sigma_{\mathrm{m}}^2/m_p^2) + 0.0068(\sigma_{\mathrm{m}}/m_p) + 1.53 \times 10^{-5}}$$

Note that m_p is a constant of the system, not of the signal. The peak signal A can vary form talker to talker, whereas m_p is fixed for a given system.

12.5-1 A message signal m(t) with

$$S_{\mathrm{m}}(\omega) = \frac{\alpha^2}{\omega^2 + \alpha^2} \qquad (\alpha = 3000\pi)$$

DSB-SC modulates a carrier of 100 kHz. Assume an ideal channel with $H_c(\omega) = 10^{-3}$ and the channel noise PSD $S_n(\omega) = 2 \times 10^{-9}$. The transmitted power is required to be 1 kW, and $G = 10^{-2}$.

(a) Determine transfer functions of optimum preemphasis and deemphasis filters.

(b) Determine the output signal power, the noise power, and the output SNR.

(c) Determine γ at the demodulator input.

12.5-2 Repeat Prob. 12.5-1 for the SSB (USB) case.

12.5-3 It was shown in the text that when the baseband m(t) is band-limited with a uniform PSD, PM and FM have identical performance from the SNR point of view. For such m(t), show that optimum PDE filters in angle modulation can improve the output SNR by a factor of 4/3 (or 1.3 dB) only. Find the optimum PDE filter transfer functions.

13

BEHAVIOR OF DIGITAL COMMUNICATION SYSTEMS IN THE PRESENCE OF NOISE

In analog systems, the chief objective is the fidelity of reproduction of waveforms, and, hence, the suitable performance criterion is the output signal-to-noise ratio. The choice of this criterion stems from the fact that the signal-to-noise ratio is related to the ability of the listener to interpret a message. In digital communication systems, the transmitter input is chosen from a finite set of possible symbols, or messages. The objective at the receiver is not to reproduce the waveform with fidelity, because the possible waveforms are already known exactly. Our goal is to decide, from the noisy received signal, which of the waveforms has been transmitted. Logically, the appropriate figure of merit in a digital communication system is the probability of error in making such a decision at the receiver.

13.1 OPTIMUM THRESHOLD DETECTION

In the threshold detection method discussed in Example 10.13, the received pulse is sampled at its peak amplitude A_p. Because of channel noise, the sampled value is not A_p but $A_p + \mathrm{n}$. The decision is made from the value $A_p + \mathrm{n}$. It was shown in Example 10.13 that for the polar binary case the error probability P_e is

$$P_e = Q(\rho) \tag{13.1}$$

where

$$\rho = \frac{A_p}{\sigma_\mathrm{n}} \tag{13.2a}$$

σ_n^2 being the variance of the received noise. To minimize P_e, we need to maximize ρ because $Q(\rho)$ decreases monotonically with ρ. Note that A_p is the signal amplitude and σ_n is the rms noise so that ρ is the signal amplitude to rms noise ratio.

Let the received pulse $p(t)$ be time limited to T_o (Fig. 13.1). Note that here we use a general symbol T_o rather than T_b because the pulse may not be binary. We shall keep the discussion as general as possible at this point. There is a possibility of increasing $\rho = A_p/\sigma_\mathrm{n}$

by passing the received pulse through a filter that enhances the pulse amplitude at some instant t_m and simultaneously reduces the noise power σ_{n}^2 (Fig. 13.1). We thus seek a filter with a transfer function $H(\omega)$ that maximizes ρ, where

$$\rho^2 = \frac{p_o^2(t_m)}{\sigma_{\mathrm{n}}^2} \tag{13.2b}$$

Because

$$\begin{aligned} p_o(t) &= \mathcal{F}^{-1}[P(\omega)H(\omega)] \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)H(\omega)e^{j\omega t}\,d\omega \end{aligned}$$

we have

$$p_o(t_m) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)H(\omega)e^{j\omega t_m}\,d\omega \tag{13.3}$$

Also,

$$\sigma_{\mathrm{n}}^2 = \overline{\mathrm{n}_o^2(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H(\omega)|^2\,d\omega \tag{13.4}$$

Hence,

$$\rho^2 = \frac{[\int_{-\infty}^{\infty} H(\omega)P(\omega)e^{j\omega t_m}\,d\omega]^2}{2\pi\int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H(\omega)|^2\,d\omega} \tag{13.5}$$

In the Schwarz inequality (Appendix B), if we identify $X(\omega) = H(\omega)\sqrt{S_{\mathrm{n}}(\omega)}$ and $Y(\omega) = P(\omega)e^{j\omega t_m}/\sqrt{S_{\mathrm{n}}(\omega)}$, then it follows from Eq. (13.5) that

$$\rho^2 \le \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|P(\omega)|^2}{S_{\mathrm{n}}(\omega)}\,d\omega \tag{13.6a}$$

with equality only if

$$H(\omega)\sqrt{S_{\mathrm{n}}(\omega)} = k\left[\frac{P(\omega)e^{j\omega t_m}}{\sqrt{S_{\mathrm{n}}(\omega)}}\right]^* = \frac{kP(-\omega)e^{-j\omega t_m}}{\sqrt{S_{\mathrm{n}}(\omega)}}$$

Hence

$$H(\omega) = k\frac{P(-\omega)e^{-j\omega t_m}}{S_{\mathrm{n}}(\omega)} \tag{13.6b}$$

where k is an arbitrary constant.

For white channel noise $S_{\mathrm{n}}(\omega) = \mathcal{N}/2$, and Eqs. (13.6) become

$$\rho_{\max}^2 = \frac{1}{\pi\mathcal{N}}\int_{-\infty}^{\infty} |P(\omega)|^2\,d\omega = \frac{2E_p}{\mathcal{N}} \tag{13.7a}$$

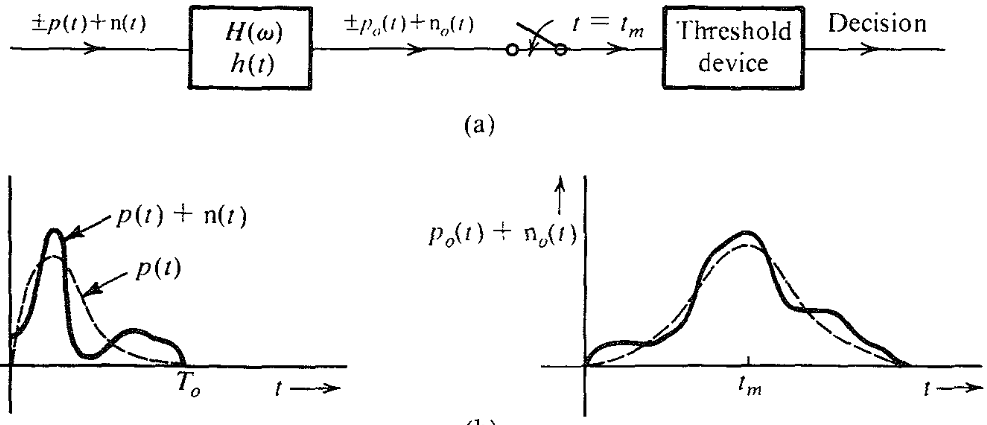
where E_p is the energy of $p(t)$, and

$$H(\omega) = k'P(-\omega)e^{-j\omega t_m} \tag{13.7b}$$

where $k' = 2k/\mathcal{N}$ is an arbitrary constant.

The unit impulse response $h(t)$ of the optimum filter is given by

$$h(t) = \mathcal{F}^{-1}[k'P(-\omega)e^{-j\omega t_m}]$$



(a)

(b)

Figure 13.1 Scheme to minimize the error probability in threshold detection.

Note that $p(-t) \Longleftrightarrow P(-\omega)$ and $e^{-j\omega t_m}$ represents the time delay of t_m seconds. Hence,

$$h(t) = k'p(t_m - t) \tag{13.7c}$$

The signal $p(t_m - t)$ is the signal $p(-t)$ delayed by t_m. Three cases, $t_m < T_o$, $t_m = T_o$, and $t_m > T_o$, are shown in Fig. 13.2. The first case, $t_m < T_o$, yields a noncausal impulse response, which is unrealizable.* Although the other two cases yield physically realizable filters, the last case, $t_m > T_o$, delays the decision-making instant t_m an unnecessary length of time. The case $t_m = T_o$ gives the minimum delay for decision making using a realizable filter. In our future discussion, we shall assume $t_m = T_o$, unless otherwise mentioned.

Observe that both $p(t)$ and $h(t)$ have a width of T_o seconds. Hence $p_o(t)$, which is a convolution of $p(t)$ and $h(t)$, has a width of $2T_o$ seconds, with its peak occurring at $t = T_o$. Also, because $P_o(\omega) = P(\omega)H(\omega) = k'|P(\omega)|^2 e^{-j\omega T_o}$, $p_o(t)$ is symmetrical about $t = T_o$ (Fig. 13.1).†

The arbitrary constant k' in Eq. (13.7) multiplies both the signal and the noise by the same factor and does not affect the ratio ρ. Hence, the error probability, or the system performance, is independent of the value of k'. For convenience, we choose $k' = 1$. This gives

$$h(t) = p(T_o - t) \tag{13.8a}$$

and

$$H(\omega) = P(-\omega)e^{-j\omega T_o} \tag{13.8b}$$

---

* The filter unrealizability can be readily understood intuitively when the decision-making instant is $t_m < T_o$. In this case, we are forced to make a decision even before the complete pulse is fed to the filter ($t_m < T_o$). This calls for a prophetic filter, which can respond to inputs even before they are applied. As we know, only unrealizable (noncausal) filters can do this job.

† This follows from the fact that because $|P(\omega)|^2$ is an even function of ω, its inverse transform is symmetrical about $t = 0$ (see Prob. 3.1-1). The output from the previous input pulse terminates and has a zero value at $t = T_o$. Similarly, the output from the following pulse starts and has a zero value at $t = T_o$. Hence, at the decision-making instant T_o, no intersymbol interference occurs.

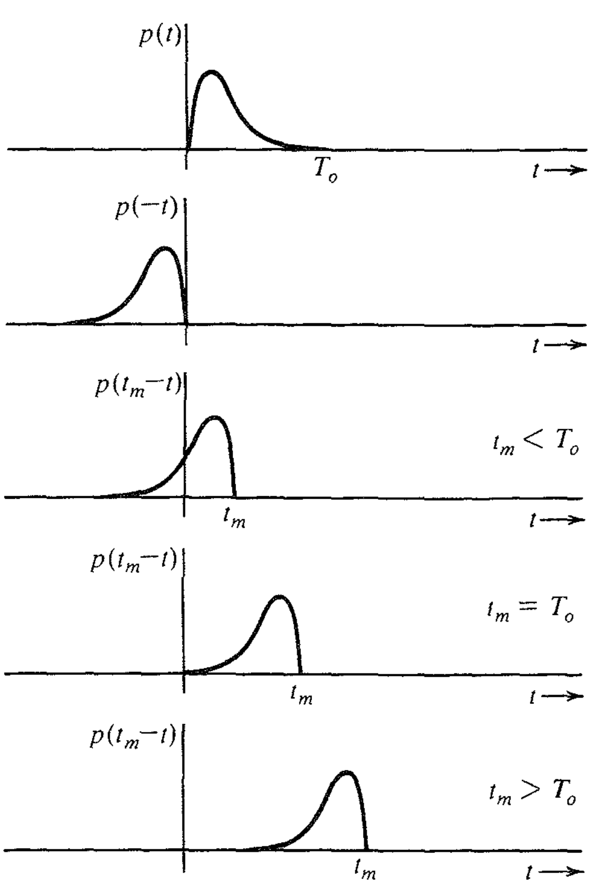


Figure 13.2 Optimum choice for sampling instant.

The optimum filter in Eqs. (13.8) is known as the **matched filter.**[*] At the output of this filter, the signal to rms noise amplitude ratio is maximum at the decision-making instant $t = T_o$.

The matched filter is optimum in the sense that it maximizes the signal amplitude to rms noise ratio at the decision-making instant. Although it is reasonable to assume that maximization of this particular signal to noise ratio will minimize the detection error probability, we have not proved that threshold detection (sample and decide) is the optimum method from the detection error point of view. It will be shown in Chapter 14 that when the channel noise is white gaussian, the matched-filter receiver is indeed the optimum receiver that minimizes the detection error probability. The maximum value of this signal to rms noise ratio attained by the matched filter is given in Eq. (13.7a). The peak amplitude $p_o(t_m) = A_p$ is found by substituting Eq. (13.8b) (with $k' = 1$) into Eq. (13.3),

$$A_p = \frac{1}{2\pi}\int_{-\infty}^{\infty} |P(\omega)|^2\,d\omega = E_p \tag{13.9a}$$

The noise power σ_n^2 is obtained by substituting Eq. (13.8b) (with $k' = 1$) into Eq. (13.4),

$$\sigma_n^2 = \frac{\mathcal{N}}{4\pi}\int_{-\infty}^{\infty} |P(\omega)|^2\,d\omega = \frac{\mathcal{N}E_p}{2} \tag{13.9b}$$

---

[*] It is important to remember that the optimum filter is the matched filter only when the channel noise is white. For a general case, the optimum filter is given in Eq. (13.6b).

Hence,

$$\rho_{\text{max}}^2 = \frac{A_p^2}{\sigma_n^2} = \frac{2E_p}{\mathcal{N}} \tag{13.9c}$$

and

$$P_e = Q(\rho_{\text{max}}) = Q\left(\sqrt{\frac{2E_p}{\mathcal{N}}}\right) \tag{13.9d}$$

Equations (13.9) are truly remarkable. They show that at the decision-making instant, the signal amplitude, and the rms noise amplitude depend on the waveform $p(t)$ only through its energy E_p. As far as the system performance is concerned, when the matched-filter receiver is used, all the waveforms used for $p(t)$ are equivalent as long as they have the same energy.

The matched filter may also be realized by the alternative arrangement shown in Fig. 13.3. If the input to the matched filter is $\mathrm{r}(t)$, then the output $\mathrm{y}(t)$ is given by

$$\mathrm{y}(t) = \int_{-\infty}^{\infty} \mathrm{r}(x)h(t-x)\,dx$$

where $h(t) = p\,(T_o - t)$ and

$$h(t-x) = p\,[T_o - (t-x)] = p\,(x + T_o - t)$$

Hence,

$$\mathrm{y}(t) = \int_{-\infty}^{\infty} \mathrm{r}(x)p\,(x + T_o - t)\,dx \tag{13.10a}$$

At the decision-making instant $t = T_o$, we have

$$\mathrm{y}(T_o) = \int_{-\infty}^{\infty} \mathrm{r}(x)\,p\,(x)\,dx \tag{13.10b}$$

Because the input $\mathrm{r}(x)$ is assumed to start at $x = 0$ and $p\,(x) = 0$ for $x > T_o$,

$$\mathrm{y}(T_o) = \int_{0}^{T_o} \mathrm{r}(x)\,p\,(x)\,dx \tag{13.10c}$$

We can implement Eqs. (13.10) as shown in Fig. 13.3. This type of arrangement is known as the **correlation receiver** and is equivalent to the matched-filter receiver.

The right-hand side of Eq. (13.10a) is $\psi_{rp}(T_o - t)$, where $\psi_{rp}(\tau)$ is the crosscorrelation of the received pulse with $p(t)$. Recall that correlation basically measures the similarity of signals (see Sec. 2.6). Thus, the optimum detector measures the similarity of the received signal with the pulse $p(t)$. Based on this similarity measure, it decides whether $p(t)$ was transmitted or not.

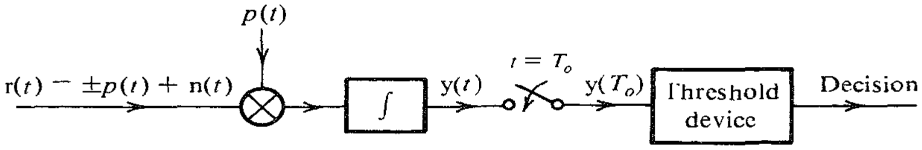


Figure 13.3 Correlation detector.

Thus far we have discussed polar signaling where only one basic pulse $p(t)$ is used. Generally, in binary communication, we use two distinct pulses $p(t)$ and $q(t)$ to represent the two symbols. The optimum receiver for such a case will now be discussed.

13.2 GENERAL ANALYSIS: OPTIMUM BINARY RECEIVER

In a binary scheme where symbols are transmitted every T_b seconds, let $p(t)$ and $q(t)$ be the two pulses used to transmit **1** and **0**. The optimum receiver structure considered here is shown in Fig. 13.4a. The incoming pulse is transmitted through a filter $H(\omega)$, and the output $\mathrm{r}(t)$ is sampled at T_b. The decision as to whether **0** or **1** was present at the input depends on whether $\mathrm{r}(T_b) <$ or $> a_o$, where a_o is the optimum threshold.

Let $p_o(t)$ and $q_o(t)$ be the response of $H(\omega)$ to inputs $p(t)$ and $q(t)$, respectively. From Eq. (13.3) it follows that

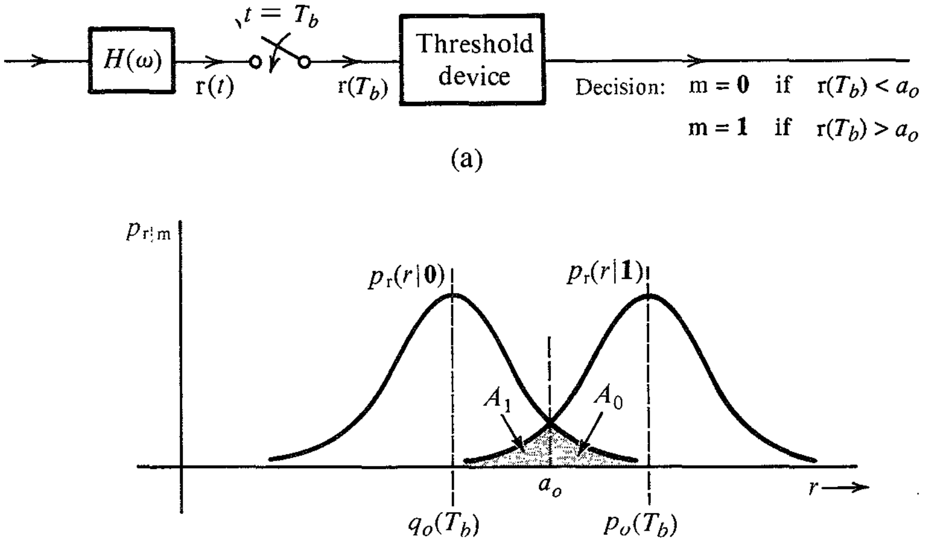
$$p_o(T_b) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)H(\omega)e^{j\omega T_b}\,d\omega \tag{13.11a}$$

$$q_o(T_b) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Q(\omega)H(\omega)e^{j\omega T_b}\,d\omega \tag{13.11b}$$

and σ_{n}^2, the variance, or power, of the noise at the filter output, is

$$\sigma_{\mathrm{n}}^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H(\omega)|^2\,d\omega \tag{13.11c}$$

If n is the noise output at T_b, then the sampler output $\mathrm{r}(T_b) = q_o(T_b) + \mathrm{n}$ or $p_o(T_b) + \mathrm{n}$, depending on whether $\mathrm{m} = \mathbf{0}$ or $\mathrm{m} = \mathbf{1}$, is received. Hence, r is a gaussian RV of variance σ_{n}^2 and mean $q_o(T_b)$ or $p_o(T_b)$, depending on whether $\mathrm{m} = \mathbf{0}$ or $\mathbf{1}$. Thus, the conditional PDFs of the sampled output $\mathrm{r}(T_b)$ are


$H(\omega)$
$t = T_b$
$\mathrm{r}(t)$
$\mathrm{r}(T_b)$
Threshold device
Decision: $\mathrm{m} = \mathbf{0}$ if $\mathrm{r}(T_b) < a_o$
$\mathrm{m} = \mathbf{1}$ if $\mathrm{r}(T_b) > a_o$
$p_{\mathrm{r}|\mathrm{m}}$
$p_{\mathrm{r}}(r|\mathbf{0})$
$p_{\mathrm{r}}(r|\mathbf{1})$
A_1
A_0
a_o
r
$q_o(T_b)$
$p_o(T_b)$

(a)

(b)

Figure 13.4 Optimum binary threshold detection.

$$p_{\mathrm{r|m}}(r|\mathbf{0}) = \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} e^{-[r-q_o(T_b)]^2/2\sigma_{\mathrm{n}}^2}$$

$$p_{\mathrm{r|m}}(r|\mathbf{1}) = \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} e^{-[r-p_o(T_b)]^2/2\sigma_{\mathrm{n}}^2}$$

The two PDFs are shown in Fig. 13.4b. If a_o is the optimum threshold of detection, then the decision is $\mathrm{m} = \mathbf{0}$ if $\mathrm{r} < a_o$ and $\mathrm{m} = \mathbf{1}$ if $\mathrm{r} > a_o$. The conditional error probability $P(\epsilon \mid \mathrm{m} = \mathbf{0})$ is the probability of making a wrong decision when $\mathrm{m} = \mathbf{0}$. This is simply the area A_0 under $p_{\mathrm{r|m}}(r|\mathbf{0})$ from a_o to ∞. Similarly, $P(\epsilon|\mathrm{m} = \mathbf{1})$ is the area A_1 under $p_{\mathrm{r|m}}(r|\mathbf{1})$ from $-\infty$ to a_o (Fig. 13.4b), and

$$P_e = \sum_i P(\epsilon|m_i)P(m_i) = \tfrac{1}{2}(A_0 + A_1)$$

assuming $P_{\mathrm{m}}(0) = P_{\mathrm{m}}(1) = 0.5$. From Fig. 13.4b it can be seen that the sum $A_0 + A_1$ of the shaded areas is minimized by choosing a_o at the intersection of the two PDFs. Thus,

$$a_o = \frac{p_o(T_b) + q_o(T_b)}{2} \tag{13.12a}$$

and the corresponding P_e is

$$\begin{aligned} P_e &= P(\epsilon|\mathbf{0}) = P(\epsilon|\mathbf{1}) \\ &= \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} \int_{a_o}^{\infty} e^{-[r-q_o(T_b)]^2/2\sigma_{\mathrm{n}}^2}\, dr \\ &= Q\left[\frac{a_o - q_o(T_b)}{\sigma_{\mathrm{n}}}\right] \\ &= Q\left[\frac{p_o(T_b) - q_o(T_b)}{2\sigma_{\mathrm{n}}}\right] && (13.12\text{b}) \\ &= Q\left(\frac{\beta}{2}\right) && (13.12\text{c}) \end{aligned}$$

where we define

$$\beta = \frac{p_o(T_b) - q_o(T_b)}{\sigma_{\mathrm{n}}} \tag{13.13}$$

Substituting Eqs. (13.11) into Eq. (13.13), we get

$$\beta^2 = \frac{\left[\int_{-\infty}^{\infty}[P(\omega) - Q(\omega)]H(\omega)e^{j\omega T_b}\, d\omega\right]^2}{2\pi \int_{-\infty}^{\infty} S_{\mathrm{n}}(\omega)|H(\omega)|^2\, d\omega}$$

This equation is of the same form as Eq. (13.5) with $P(\omega)$ replaced by $P(\omega) - Q(\omega)$. Hence,

$$\beta_{\max}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|P(\omega) - Q(\omega)|^2}{S_{\mathrm{n}}(\omega)}\, d\omega \tag{13.14a}$$

and the optimum filter $H(\omega)$ is given by

$$H(\omega) = k\frac{[P(-\omega) - Q(-\omega)]e^{-j\omega T_b}}{S_n(\omega)} \tag{13.14b}$$

where k is an arbitrary constant.

For white noise $S_n(\omega) = \mathcal{N}/2$, and the optimum filter $H(\omega)$ is given by*

$$H(\omega) = [P(-\omega) - Q(-\omega)]e^{-j\omega T_b} \tag{13.15a}$$

and

$$h(t) = p\,(T_b - t) - q\,(T_b - t) \tag{13.15b}$$

This is a filter matched to the pulse $p(t) - q(t)$. The corresponding β is [Eq. (13.14a)]

$$\beta^2_{\max} = \frac{1}{\pi\mathcal{N}}\int_{-\infty}^{\infty} |P(\omega) - Q(\omega)|^2\,d\omega \tag{13.16a}$$

$$= \frac{2}{\mathcal{N}}\int_0^{T_b} [p(t) - q(t)]^2\,dt \tag{13.16b}$$

$$= \frac{E_p + E_q - 2E_{pq}}{\mathcal{N}/2} \tag{13.16c}$$

where E_p and E_q are the energies of $p(t)$ and $q(t)$, respectively, and

$$E_{pq} = \int_0^{T_b} p(t)q(t)\,dt \tag{13.17}$$

So far, we have been using the notation P_e to denote error probability. In the binary case, this error probability is the **bit error probability** or **bit error rate** (BER), and will be denoted by P_b (rather than P_e). Thus, from Eqs. (13.12c) and (13.16c),

$$P_b = Q\left(\frac{\beta_{\max}}{2}\right) \tag{13.18a}$$

$$= Q\left(\sqrt{\frac{E_p + E_q - 2E_{pq}}{2\mathcal{N}}}\right) \tag{13.18b}$$

The optimum threshold a_o is obtained by substituting Eqs. (13.11a, b) and (13.15a) into Eq. (13.12a) and recognizing that (see Prob. 3.7-3)

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} P\,(\omega)Q\,(-\omega)\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} P\,(-\omega)Q\,(\omega)\,d\omega = E_{pq}$$

This gives

$$a_o = \tfrac{1}{2}(E_p - E_q) \tag{13.19}$$

In deriving the optimum binary receiver, we assumed a certain receiver structure (the threshold-detection receiver in Fig. 13.4). It is not clear yet whether there exists another structure that may have better performance than that in Fig. 13.4. It will be shown in Chapter 14 that for a gaussian noise, the receiver derived here is the absolute optimum. Equation (13.18b) gives P_b

---

* Because k in Eq. (13.14b) is arbitrary, we choose $k = \mathcal{N}/2$ for convenience.

for the optimum receiver when the channel noise is white. For the case of nonwhite noise, P_b is obtained by substituting $\beta_{\max}$ from Eq. (13.14a) into Eq. (13.18a).

13.2.1 Equivalent Optimum Binary Receivers

For the optimum receiver in Fig. 13.4a,

$$H(\omega) = P(-\omega)e^{-j\omega T_b} - Q(-\omega)e^{-j\omega T_b}$$

This filter can be realized as a parallel combination of two filters matched to $p(t)$ and $q(t)$, respectively, as shown in Fig. 13.5a. Yet another equivalent form is shown in Fig. 13.5b. Because the threshold is $(E_p - E_q)/2$, we subtract $E_p/2$ and $E_q/2$, respectively, from the two matched filter outputs. This is equivalent to shifting the threshold to 0. In the case where $E_p = E_q$, we need not subtract $E_p/2$ and $E_q/2$ from the two outputs, and the receiver simplifies to that shown in Fig. 13.5c.

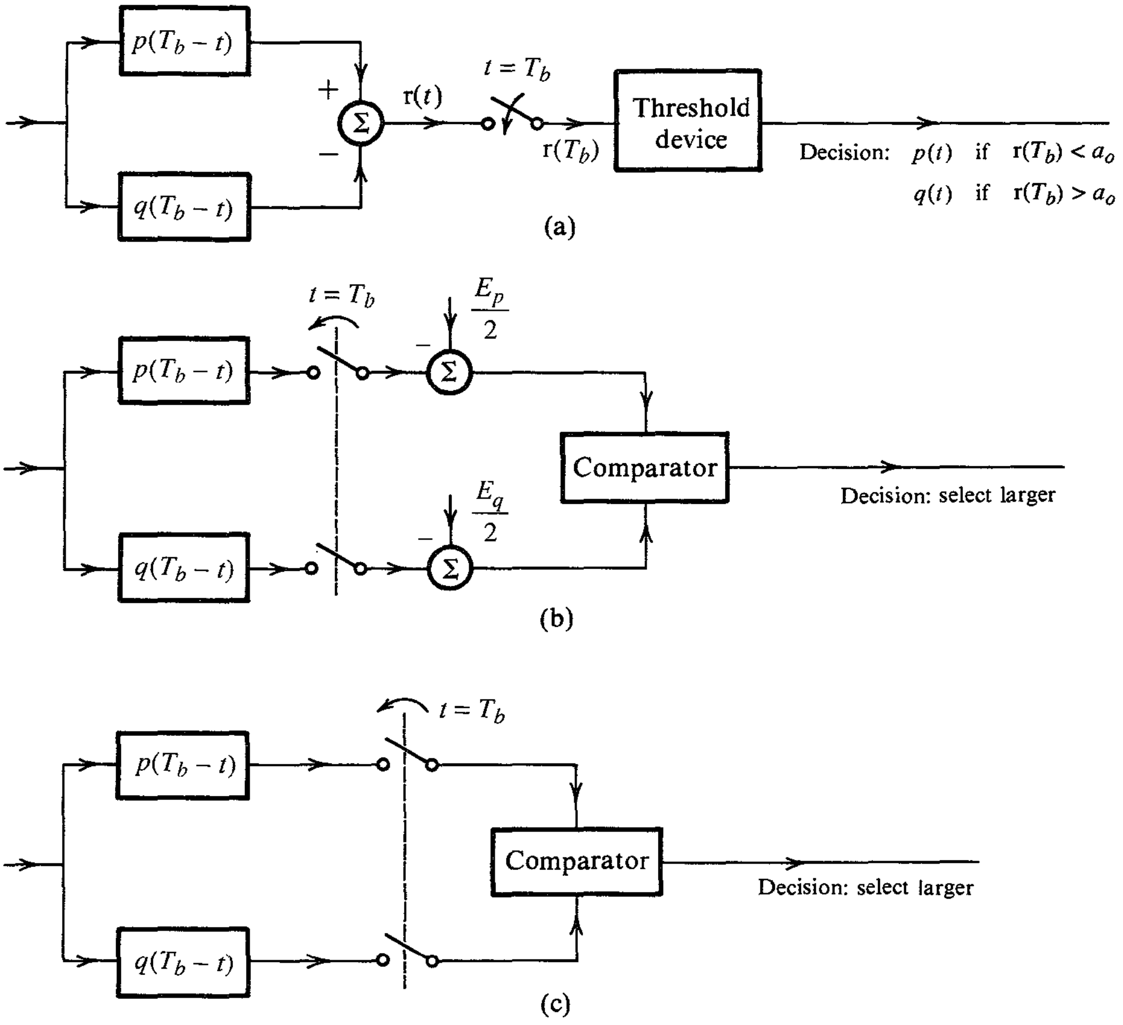

$p(T_b - t)$
+
Σ
−
r(t)
$t = T_b$
r(T_b)
$q(T_b - t)$
Threshold device
Decision: $p(t)$ if r(T_b) < a_o
$q(t)$ if r(T_b) > a_o
(a)
$t = T_b$
$p(T_b - t)$
$\frac{E_p}{2}$
−
Σ
Comparator
Decision: select larger
$\frac{E_q}{2}$
−
Σ
$q(T_b - t)$
(b)
$t = T_b$
$p(T_b - t)$
Comparator
Decision: select larger
$q(T_b - t)$
(c)

Figure 13.5 Realization of the optimum binary threshold detector.

Polar Signaling

In this case $q(t) = -p(t)$. Hence,

$$E_p = E_q \qquad \text{and} \qquad E_{pq} = -\int_{-\infty}^{\infty} p^2(t)\,dt = -E_p$$

Substituting these results into Eq. (13.18b) yields

$$P_b = Q\left(\sqrt{\frac{2E_p}{\mathcal{N}}}\right) \tag{13.20a}$$

Also from Eq. (13.15b),

$$h(t) = 2p(T_b - t)$$

Recall that the multiplication of $h(t)$ by any constant multiplies both the signal and the noise by the same factor, and hence does not affect the system performance. For convenience, we shall multiply $h(t)$ by 0.5 to obtain

$$h(t) = p(T_b - t) \tag{13.20b}$$

From Eq. (13.19), the threshold a_o is

$$a_o = 0 \tag{13.20c}$$

Therefore, for the polar case, the receiver in Fig. 13.5a reduces to that shown in Fig. 13.6a with threshold 0. This filter is equivalent to that in Fig. 13.3.

The error probability can be expressed in terms of a more basic parameter E_b, the energy per bit:

$$E_b = \text{Energy per bit}$$

In the polar case assuming **1** and **0** are equally likely, the bit energy E_b is the mean of E_p and E_q. Hence,

$$E_b = \frac{E_p + E_q}{2} = E_p \tag{13.21a}$$

and from Eq. (13.20a),

$$P_b = Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right) \tag{13.21b}$$

The parameter $E_b/\mathcal{N}$ is the normalized energy per bit, which will be seen in future discussions as a fundamental parameter serving as a figure of merit in digital communication.* Because the signal power is equal to E_b times the bit rate, a given E_b is the same as a given signal power (for a given bit rate). Hence, when we compare a system for a given value of E_b, we are comparing it for a given signal power. Using an asymptotic approximation [Eq. (10.36a)] for $Q(\sqrt{2E_b/\mathcal{N}})$, we obtain

---

* If the transmission rate is R_b pulses per second, the signal power S_i is $S_i = E_b R_b$, and $E_b/\mathcal{N} = S_i/\mathcal{N}R_b$. Observe that $S_i/\mathcal{N}R_b$ is similar to the parameter γ (signal-to-noise ratio $S_i/\mathcal{N}B$) used in analog systems.

$$P_b \simeq \frac{1}{2\sqrt{\pi E_b/\mathcal{N}}} e^{-E_b/\mathcal{N}} \qquad E_b/\mathcal{N} \gg 1 \tag{13.21c}$$

Figure 13.6b shows the plot of P_b as a function of $E_b/\mathcal{N}$ (in decibels). Equation (13.21b) indicates that, for optimum threshold detection, the system performance does not depend on the pulse shape, but on the pulse energy.

On-Off Signaling

In this case, $q(t) = 0$, and Fig. 13.5a reduces to Fig. 13.6a (the same as that for the polar case) except that the threshold, as seen from Eq. (13.19), is $E_p/2$. Because $q(t) = 0$, From Eqs. (13.17) and (13.18),

$$E_q = 0, \qquad E_{pq} = 0, \qquad \text{and} \qquad P_b = Q\left(\sqrt{\frac{E_p}{2\mathcal{N}}}\right)$$

Also

$$E_b = \frac{E_p + E_q}{2} = \frac{E_p}{2}$$

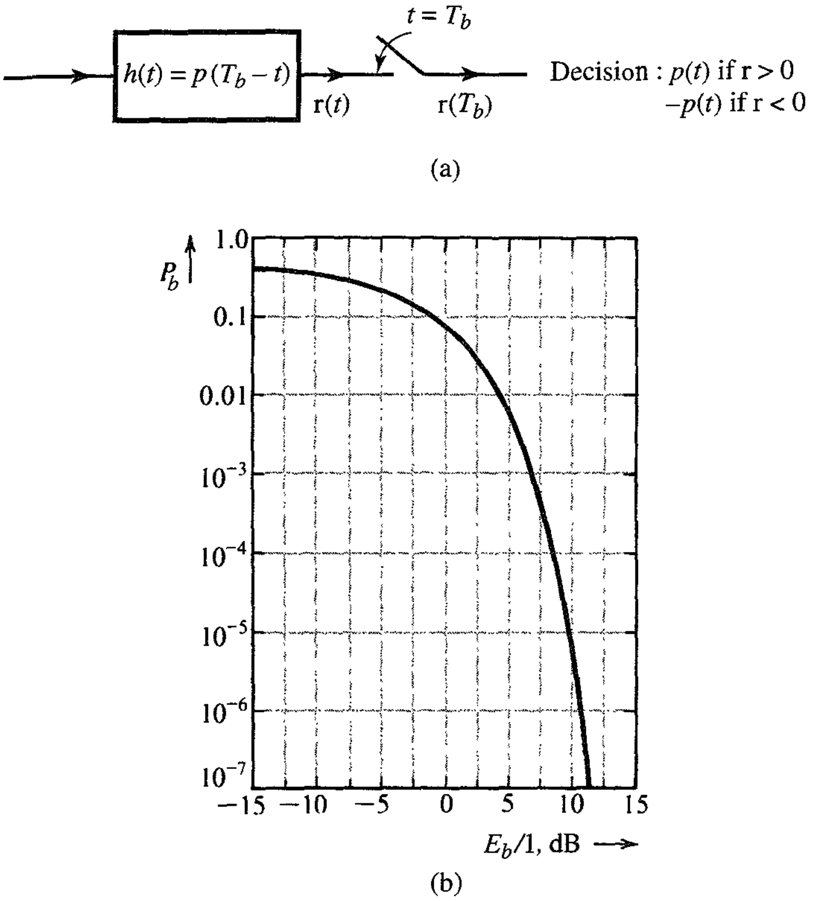


Figure 13.6 Optimum threshold detector and its error probability for polar signaling.

Therefore,

$$P_b = Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right) \tag{13.22a}$$

$$\simeq \frac{1}{\sqrt{2\pi E_b/\mathcal{N}}} e^{-E_b/2\mathcal{N}} \qquad E_b/\mathcal{N} \gg 1 \tag{13.22b}$$

A comparison of this with Eq. (13.21b) shows that on-off signaling requires twice as much energy per bit (3 dB more power) to achieve the same performance (i.e., the same P_b) as polar signaling.

Orthogonal Signaling

In orthogonal signaling, $p(t)$ and $q(t)$ are selected to be orthogonal over the interval $(0, T_b)$. This gives

$$E_{pq} = \int_0^{T_b} p(t)q(t)\,dt = 0$$

Two examples of binary orthogonal pulses are shown in Fig. 13.7. From Eq. (13.18),

$$P_b = Q\left(\sqrt{\frac{E_p + E_q}{2\mathcal{N}}}\right) \tag{13.23}$$

Assuming **1** and **0** to be equiprobable,

$$E_b = \frac{E_p + E_q}{2}$$

and

$$P_b = Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right) \tag{13.24a}$$

$$\simeq \frac{1}{\sqrt{2\pi E_b/\mathcal{N}}} e^{-E_b/2\mathcal{N}} \qquad E_b/\mathcal{N} \gg 1 \tag{13.24b}$$

This shows that the performance of orthogonal signaling is inferior to that of polar signaling by 3 dB, but it is identical to that of on-off signaling. This is not surprising in view of the fact that on-off *is* a special case of orthogonal signaling because $E_{pq} = 0$.

Bipolar Signaling

Bipolar signaling, although a binary scheme, uses three symbols: $p(t)$, $-p(t)$, and 0. Hence, the preceding results cannot be used directly without some modifications. In this case, we seek to distinguish between $p(t)$ and 0 or $-p(t)$ and 0 so that it is basically similar to the on-off case. Hence, we can use the receiver in Fig. 13.6a with threshold $\pm a_o = \pm E_p/2$. Thus, if $|\mathrm{r}(T_b)| < E_p/2$, the decision is **0**, and if $|\mathrm{r}(T_b)| > E_p/2$, the decision is **1**. This receiver is a matched filter in Fig. 13.1a. To compute P_b, we compute the conditional error probabilities. When **0** is transmitted by no pulse, the receiver output is just the noise n with variance $\sigma_\mathrm{n}^2 = \mathcal{N}E_p/2$ [see Eq. (13.9b)]. Hence,

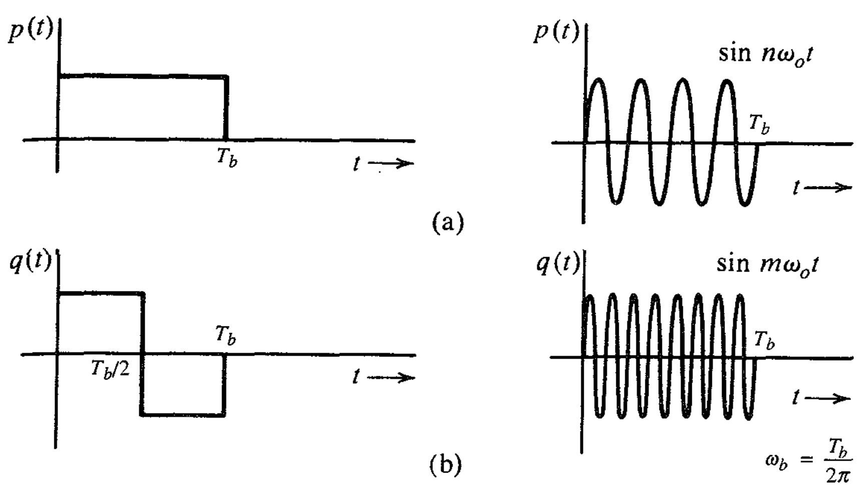


Figure 13.7 Examples of orthogonal signals.

$$
\begin{aligned}
P(\epsilon|\mathbf{0}) &= \text{probability}\left(|n| > \frac{E_p}{2}\right)\\
&= 2Q\left(\frac{E_p}{2\sigma_n}\right)\\
&= 2Q\left(\sqrt{\frac{E_p}{2\mathcal{N}}}\right)
\end{aligned}
$$

When **1** is transmitted, the filter output at T_b is $A_p + \mathrm{n}$, where $A_p = E_p$ when $p(t)$ is transmitted [Eq. (13.9a)]. Consequently, $A_p = -E_p$ when $-p(t)$ is transmitted. Thus,

$$
\begin{aligned}
P(\epsilon|\mathbf{1}) &= \text{probability}\left(n < -\frac{E_p}{2}\right) \quad \text{when } p(t) \text{ is used, or}\\
&\quad\ \text{probability}\left(n > \frac{E_p}{2}\right) \quad \text{when } -p(t) \text{ is used}\\
&= Q\left(\frac{E_p}{2\sigma_n}\right)\\
&= Q\left(\sqrt{\frac{E_p}{2\mathcal{N}}}\right)
\end{aligned}
$$

Therefore, the average error probability is (assuming **1** and **0** equally likely)

$$
P_b = \frac{1}{2}[P(\epsilon|\mathbf{0}) + P(\epsilon|\mathbf{1})] = 1.5Q\left(\sqrt{\frac{E_p}{2\mathcal{N}}}\right)
$$

This shows that P_b is 50% higher for bipolar signaling than for on-off signaling. But this difference can be compensated by a very small increase in E_p because the Q function, which varies exponentially with E_p, is extremely sensitive to variations in E_p. For example, to realize $P_b = 0.286 \times 10^{-6}$, we need $(E_p)_{\text{bipolar}} = 1.016(E_p)_{\text{on-off}}$.

13.3 CARRIER SYSTEMS: ASK, FSK, PSK, AND DPSK

In digital carrier systems, baseband pulses modulate a high-frequency carrier. We have briefly discussed amplitude-shift keying (ASK), frequency-shift keying (FSK), and phase-shift keying (PSK) in Chapter 7.

Figure 13.8 shows the three schemes, using a rectangular baseband pulse. The baseband pulse may be specifically shaped (e.g., a raised cosine) to eliminate intersymbol interference and to have a finite bandwidth.

The error probability of the optimum detector depends only on the pulse energy, not on the pulse shape. Hence, as far as the error probability is concerned, the performance of a modulated scheme will be identical to that of the baseband scheme of the same energy.

The incoming modulated pulses can be demodulated either coherently (synchronously) or noncoherently (by envelope detection). The former method is the optimum and requires much more sophisticated equipment. Naturally, it has a superior performance in comparison to the latter method.

13.3.1 Coherent Detection

Let the RF pulse $p(t) = \sqrt{2}\, p'(t) \cos \omega_c t$, where $p'(t)$ is a baseband pulse. The RF pulse can be detected by a filter matched to the RF pulse $p(t)$ followed by a sampler (Fig. 13.9a). In this case [Eq. (13.7a)],

$$\rho^2 = \frac{2E_p}{\mathcal{N}}$$

where E_p is the energy of $p(t)$.

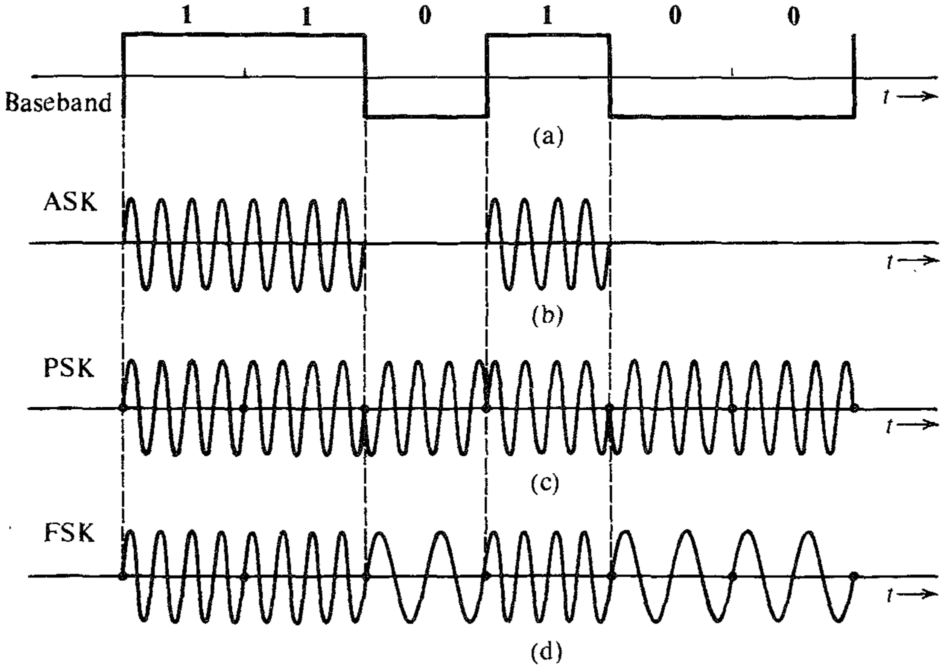


Figure 13.8 Digital modulated waveforms.

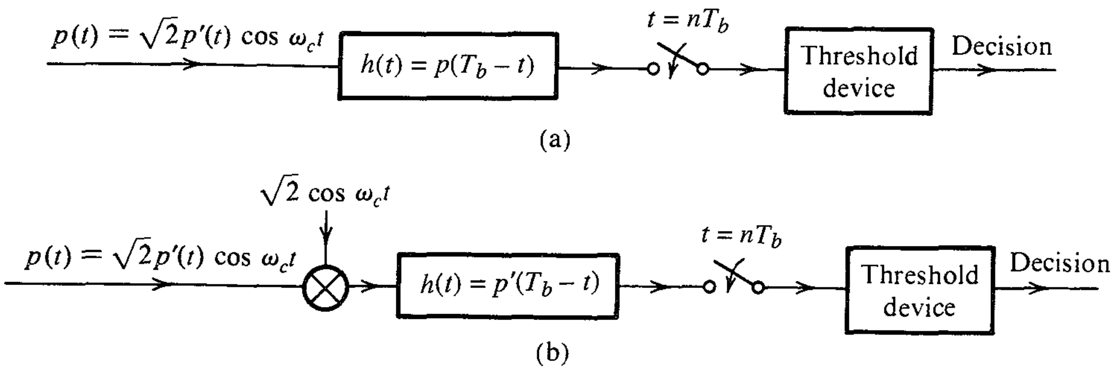


Figure 13.9 Coherent detection of digital modulated signals.

We may also detect the RF pulse by first demodulating it coherently by multiplying it by $\sqrt{2}\cos\omega_c t$. The product is the baseband pulse* $p'(t)$ plus a baseband noise with PSD $\mathcal{N}/2$ (see Example 11.11 and Sec. 12.2), and this is applied to a filter matched to the baseband pulse $p'(t)$. Because $p(t) = \sqrt{2}\,p'(t)\cos\omega_c t$, $E_p = E_{p'}$, and in this case also

$$\rho^2 = \frac{2E_p}{\mathcal{N}}$$

Hence, the two schemes are equivalent.

Let us consider the cases of PSK, ASK, and FSK individually.

Phase-Shift Keying

This is a case of polar signaling, and the results derived earlier [Eqs. (13.21b, c)] apply. The optimum detector is shown in Fig. 13.9 with threshold 0. From Eq. (13.21b),

$$P_b = Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right) \tag{13.25a}$$

$$\simeq \frac{1}{2\sqrt{\pi E_b/\mathcal{N}}}e^{-E_b/\mathcal{N}} \qquad E_b/\mathcal{N} \gg 1 \tag{13.25b}$$

Amplitude-Shift Keying

This is a case of on-off signaling, and Eqs. (13.22) apply. Recall that the optimum detector for on-off signaling is the same as that for polar. Hence, the optimum detector for ASK is the same as that for PSK with threshold $E_p/2$ (Fig. 13.9). From Eqs. (13.22) we have

$$P_b = Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right) \tag{13.26a}$$

$$\simeq \frac{1}{\sqrt{2\pi E_b/\mathcal{N}}}e^{-E_b/\mathcal{N}/2} \qquad E_b/\mathcal{N} \gg 1 \tag{13.26b}$$

Comparison of Eqs. (13.26) with Eqs. (13.25) shows that for the same performance, the pulse energy in ASK must be twice that in PSK. Hence, ASK requires 3 dB more power than PSK. Thus, in coherent detection, PSK is always preferable to ASK. For this reason, ASK is of no practical importance in coherent detection. But ASK can be useful in noncoherent (envelope) detection. In PSK, the information lies in the phase, and, hence, it cannot be detected noncoherently.

---

* There is also a spectrum of $p'(t)$ centered at $2\omega_c$, which is eventually eliminated by the filter matched to $p'(t)$.

The baseband pulses used in carrier systems should be shaped to minimize the ISI. The bandwidth of the PSK or ASK signal is twice that of the corresponding baseband signal because of modulation.*

Frequency-Shift Keying

In FSK, binary **0** and **1** are transmitted by RF pulses $\sqrt{2}p'(t)\cos\,[\omega_c - (\Delta\omega/2)]t$ and $\sqrt{2}p'(t)\cos\,[\omega_c + (\Delta\omega/2)]t$, respectively. Such a waveform may be considered to be two interleaved ASK waves. Hence, the PSD will consist of two PSDs, centered at $[f_c - (\Delta f/2)]$ and $[f_c + (\Delta f/2)]$. For a large $\Delta f/f_c$, the PSD will consist of two nonoverlapping PSDs. For a small $\Delta f/f_c$, the two spectra merge, and the bandwidth decreases. But in no case is the bandwidth less than that of ASK or PSK.

The receiver in Fig. 13.5a or b can serve as the optimum receiver. But because the pulses have equal energy, the simplified form of the optimum receiver in Fig. 13.5c is the most convenient. The filters $p\,(T_b - t)$ and $q\,(T_b - t)$ are matched to the two RF pulses and can be replaced by respective synchronous demodulators followed by filters matched to the baseband pulse $p'(t)$.

Consider the case:

$$q(t) = \sqrt{2}\,A\,\cos\left(\omega_c - \frac{\Delta\omega}{2}\right)t$$

$$p(t) = \sqrt{2}\,A\,\cos\left(\omega_c - \frac{\Delta\omega}{2}\right)t$$

To compute P_b from Eq. (13.18b), we need E_{pq},

$$\begin{aligned}
E_{pq} &= \int_0^{T_b} p(t)q(t)\,dt \\
&= 2A^2 \int_0^{T_b} \cos\left(\omega_c - \frac{\Delta\omega}{2}\right)t\,\cos\left(\omega_c + \frac{\Delta\omega}{2}\right)t\,dt \\
&= A^2\left[\int_0^{T_b} \cos\,(\Delta\omega)t\,dt + \int_0^{T_b} \cos 2\omega_c t\,dt\right] \\
&= A^2 T_b\left[\frac{\sin\,(\Delta\omega)T_b}{(\Delta\omega)T_b} + \frac{\sin\,2\omega_c T_b}{2\omega_c T_b}\right]
\end{aligned}$$

In practice $\omega_c T_b \gg 1$, and the second term on the right-hand side can be ignored. Therefore,

$$E_{pq} = A^2 T_b\,\operatorname{sinc}\,(\Delta\omega)T_b$$

Figure 13.10a shows sinc $(\Delta\omega)T_b$ as a function of $(\Delta\omega)T_b$.

To minimize P_b [Eq. (13.18b)], E_{pq} must be minimized. From Fig. 13.10a, the minimum value of E_{pq} is $-0.217A^2T_b$ and occurs at $(\Delta\omega)T_b = 1.43\pi$ or when

$$\Delta f = \frac{0.715}{T_b} = 0.715R_b$$

---

* We can also use QAM (quadrature multiplexing) to double the data rate.

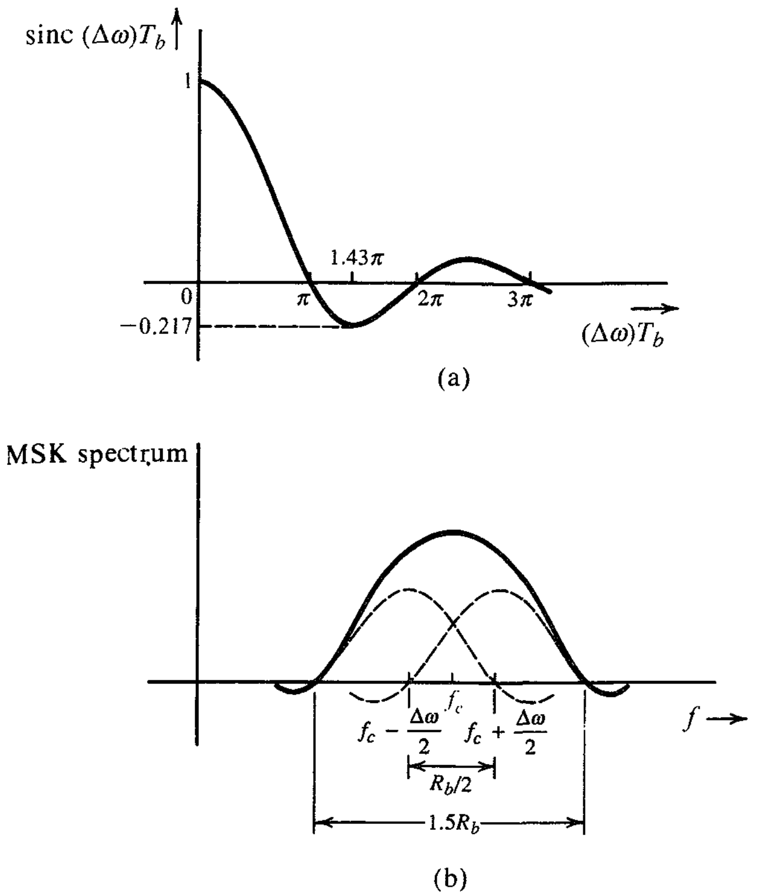


Figure 13.10 FSK and MSK spectra.

Assuming **1** and **0** equiprobable, $E_b = E_p = E_q = A^2T_b$ and $E_{pq} = -0.217A^2T_b$,

$$P_b = Q\left(\sqrt{\frac{1.217\ A^2T_b}{\mathcal{N}}}\right)$$

$$= Q\left(\sqrt{\frac{1.217E_b}{\mathcal{N}}}\right) \tag{13.27a}$$

When $E_{pq} = 0$, we have the case of orthogonal signaling. From Fig. 13.10a, it is clear that $E_{pq} = 0$ for $\Delta f = n/2T_b$, where n is any integer. Larger Δf means wider separation between signaling frequencies $\omega_c - (\Delta\omega/2)$ and $\omega_c + (\Delta\omega/2)$, and consequently larger transmission bandwidth. To minimize the bandwidth, Δf should be as small as possible. The minimum value of Δf that can be used for orthogonal signaling is $1/2T_b$. FSK using this value of Δf is known as **minimum-shift keying (MSK)** or **fast-frequency-shift keying**.

In MSK, abrupt phase changes at the bit-transition instants, characteristic of other FSK implementations, are avoided. FSK schemes where phase continuity is maintained are known as **continuous-phase FSK (CP-FSK)**, of which MSK is one example. These schemes have rapid spectral roll-off and improved efficiency.

To maintain phase continuity in CP-FSK (or MSK), the phase at every bit transition is made dependent on the past data sequence. Consider, for example, the data sequence **1001**... starting at $t = 0$. The first pulse corresponding to the first bit **1** is cos $[\omega_c + (\Delta\omega/2)]t$ over the interval 0 to T_b seconds. At $t = T_b$, this pulse ends with a phase $[\omega_c + (\Delta\omega/2)]T_b$. The next pulse, corresponding to the second data bit **0**, is $\cos\,[\omega_c - (\Delta\omega/2)]t$. This pulse is given additional phase $[\omega_c + (\Delta\omega/2)]T_b$ in order to maintain phase continuity at the transition instant. We continue this way at each transition.

MSK being an orthogonal scheme, its error probability is given by

$$P_b = Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right) \tag{13.27b}$$

This performance appears inferior to that of the optimum case in Eq. (13.27a). Closer examination shows that MSK is actually superior to the so-called optimum case. If MSK were coherently detected as ordinary FSK using an observation interval of T_b, MSK would have $P_b = Q(\sqrt{E_b/\mathcal{N}})$. But recall that MSK is CP-FSK, where the phase of each pulse is dependent on the past data sequence. Hence, better performance may be obtained by observing the received waveform over a period longer than T_b. It can be shown that if MSK is detected using an observation interval of $2T_b$, the performance of MSK is identical to that of PSK, that is,

$$P_b = Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right) \tag{13.27c}$$

MSK also has other useful properties. It has self-synchronization capabilities and its bandwidth is only $1.5R_b$, as shown in Fig. 13.10b. This is only 50% higher than for duobinary signaling. Moreover, the MSK spectrum decays much more rapidly as $1/f^4$, in contrast to the PSK (or bipolar) spectrum, which decays only as $1/f^2$ [see Eqs. (7.14) and (7.21)]. Because of these properties, MSK has received a great deal of attention recently. More discussion of MSK can be found in the references.[1,2]

13.3.2 Noncoherent Detection

If the phase Θ in the received RF pulse $\sqrt{2}p'(t)\cos\,(\omega_c t + \Theta)$ is unknown, we cannot use coherent detection techniques but must rely on noncoherent techniques, such as envelope detection. It can be shown[3,4] that when the phase Θ of the received pulse is random and uniformly distributed over $(0,\ 2\pi)$, the optimum detector is a filter matched to the RF pulse $\sqrt{2}p'(t)\cos\omega_c t$ followed by an envelope detector, a sampler (to sample at $t = T_b$), and a comparator to make the decision (Fig. 13.11).

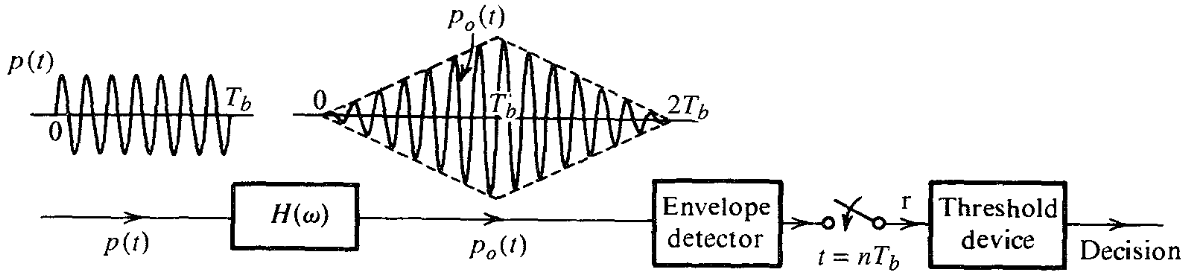


Figure 13.11 Noncoherent detection of digital modulated signals.

Amplitude-Shift Keying

The noncoherent detector for ASK is shown in Fig. 13.11. The filter $H(\omega)$ is a filter matched to the RF pulse, ignoring the phase. This means the filter output amplitude A_p will not necessarily be maximum at the sampling instant. But the envelope will be close to maximum at the sampling instant (Fig. 13.11). The matched filter output is now detected by an envelope detector. The envelope is sampled at $t = T_b$ for making the decision.

When a **1** is transmitted, the output of the envelope detector at $t = T_b$ is an envelope of a sine wave of amplitude A_p in a gaussian noise of variance σ_n^2. In this case, the envelope r has a rician density, given by [Eq. (11.64a)]

$$p_r(r|\mathrm{m}=\mathbf{1}) = \frac{r}{\sigma_n^2} e^{-(r^2+A_p^2)/2\sigma_n^2} I_0\left(\frac{rA_p}{\sigma_n^2}\right) \tag{13.28a}$$

Also, when $A_p \gg \sigma_n$ (small-noise case) from Eq. (11.64c), we have

$$p_r(r|\mathrm{m}=\mathbf{1}) \simeq \sqrt{\frac{r}{2\pi A_p \sigma_n^2}}\, e^{-(r-A_p)^2/2\sigma_n^2} \tag{13.28b}$$

$$\simeq \frac{1}{\sigma_n\sqrt{2\pi}} e^{-(r-A_p)^2/2\sigma_n^2} \tag{13.28c}$$

Observe that for small noise, the PDF of r is practically gaussian, with mean A_p and variance σ_n^2. When **0** is transmitted, the output of the envelope detector is an envelope of a gaussian noise of variance σ_n^2. The envelope in this case has a Rayleigh density, given by [Eq. (11.59)]

$$p_r(r|\mathrm{m}=\mathbf{0}) = \frac{r}{\sigma_n^2} e^{-r^2/2\sigma_n^2}$$

Both $p_r(r|\mathrm{m}=\mathbf{1})$ and $p_r(r|\mathrm{m}=\mathbf{0})$ are shown in Fig. 13.12. Using the argument used earlier (see Fig. 13.4), the optimum threshold is found to be the point where the two densities intersect. Hence, the optimum threshold a_o is

$$\frac{a_o}{\sigma_n^2} e^{-(a_o^2+A_p^2)/2\sigma_n^2} I_0\left(\frac{A_p a_o}{\sigma_n^2}\right) = \frac{a_o}{\sigma_n^2} e^{-a_o^2/2\sigma_n^2}$$

or

$$e^{-A_p^2/2\sigma_n^2} I_0\left(\frac{A_p a_o}{\sigma_n^2}\right) = 1$$

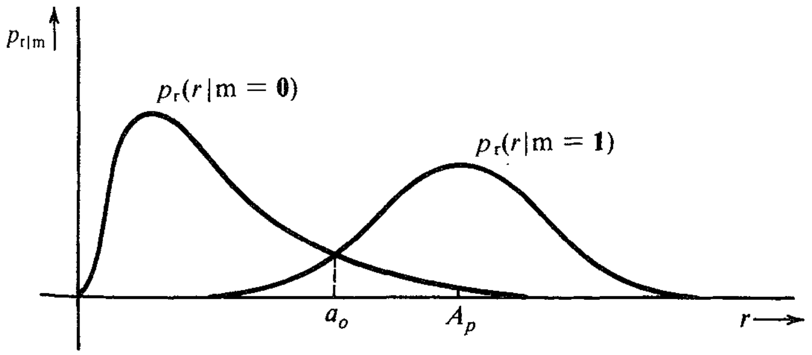


Figure 13.12 Conditional PDFs in noncoherent detection of ASK signals.

This equation is satisfied to a close approximation for

$$a_o = \frac{A_p}{2}\sqrt{1 + \frac{8\sigma_{\text{n}}^2}{A_p^2}}$$

Because the matched filter is used, $A_p = E_p$ and $\sigma_{\text{n}}^2 = \mathcal{N}E_p/2$ [see Eqs. (13.9a, b)]. Moreover, for ASK there are, on the average, only $R_b/2$ pulses per second. Thus, $E_b = E_p/2$. Hence,

$$\left(\frac{A_p}{\sigma_{\text{n}}}\right)^2 = \frac{2E_p}{\mathcal{N}} = 4\frac{E_b}{\mathcal{N}}$$

and

$$a_o = E_b\sqrt{1 + \frac{2}{E_b/\mathcal{N}}} \tag{13.29a}$$

Observe that the optimum threshold is not constant but depends on $E_b/\mathcal{N}$. This is a serious drawback in a fading channel. For a strong signal, $E_b/\mathcal{N} \gg 1$,

$$a_o \simeq E_b \tag{13.29b}$$

and

$$\begin{aligned} P(\epsilon|\text{m} = \mathbf{0}) &= \int_{A_p/2}^{\infty} p_{\text{r}}(r|\text{m} = \mathbf{0})\,dr \\ &= \int_{A_p/2}^{\infty} \frac{r}{\sigma_{\text{n}}^2} e^{-r^2/2\sigma_{\text{n}}^2}\,dr \\ &= e^{-A_p^2/8\sigma_{\text{n}}^2} \\ &= e^{-\frac{1}{2}E_b/\mathcal{N}} \end{aligned} \tag{13.30}$$

Also,

$$P(\epsilon|\text{m} = \mathbf{1}) = \int_{-\infty}^{A_p/2} p_{\text{r}}(r|\text{m} = \mathbf{1})\,dr$$

Evaluation of this integral is somewhat cumbersome.[5] For a strong signal (that is, for $E_b/\mathcal{N} \gg 1$), the rician PDF can be approximated by the gaussian PDF [Eq. (11.64c)], and

$$\begin{aligned} P(\epsilon|\text{m} = \mathbf{1}) &= \frac{1}{\sigma_{\text{n}}\sqrt{2\pi}} \int_{-\infty}^{A_p/2} e^{-(r-A_p)^2/2\sigma_{\text{n}}^2}\,dr \\ &= Q\left(\frac{A_p}{2\sigma_{\text{n}}}\right) \\ &= Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right) \end{aligned} \tag{13.31}$$

Hence,

$$P_b = P_{\text{m}}(\mathbf{0})(\epsilon|\text{m} = \mathbf{0}) + P_{\text{m}}(\mathbf{1})P(\epsilon|\text{m} = \mathbf{1})$$

Assuming $P_m(\mathbf{1}) = P_m(\mathbf{0}) = 0.5$,

$$P_b = \frac{1}{2}\left[e^{-\frac{1}{2}E_b/\mathcal{N}} + Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right)\right] \tag{13.32a}$$

Using the approximation in Eq. (10.36a),

$$P_b \simeq \frac{1}{2}\left(1 + \frac{1}{\sqrt{2\pi E_b/\mathcal{N}}}\right)e^{-\frac{1}{2}E_b/\mathcal{N}} \qquad E_b/\mathcal{N} \gg 1 \tag{13.32b}$$

$$\simeq \frac{1}{2}e^{-\frac{1}{2}E_b/\mathcal{N}} \tag{13.32c}$$

Note that in an optimum receiver, for $E_b/\mathcal{N} \gg 1$, $P(\epsilon|\mathrm{m} = \mathbf{1})$ is much smaller than $P(\epsilon|\mathrm{m} = \mathbf{0})$. For example, at $E_b/\mathcal{N} = 10$, $P(\epsilon|\mathrm{m} = \mathbf{0}) \simeq 8.7\,P(\epsilon|\mathrm{m} = \mathbf{1})$. Hence, mistaking **0** for **1** is the type of error that predominates. The timing information in noncoherent detection is extracted from the envelope of the received signal by methods discussed in Sec. 7.5.

For a coherent detector,

$$P_b = Q\left(\sqrt{\frac{E_b}{\mathcal{N}}}\right)$$

$$\simeq \frac{1}{\sqrt{2\pi E_b/\mathcal{N}}}e^{-\frac{1}{2}E_b/\mathcal{N}} \qquad E_b/\mathcal{N} \gg 1 \tag{13.33}$$

This appears similar to Eq. (13.32c) (the noncoherent case). Thus for a large $E_b/\mathcal{N}$, the performances of the coherent detector and the envelope detector are similar (Fig. 13.13). This is similar to the behavior observed in the case of analog signals.

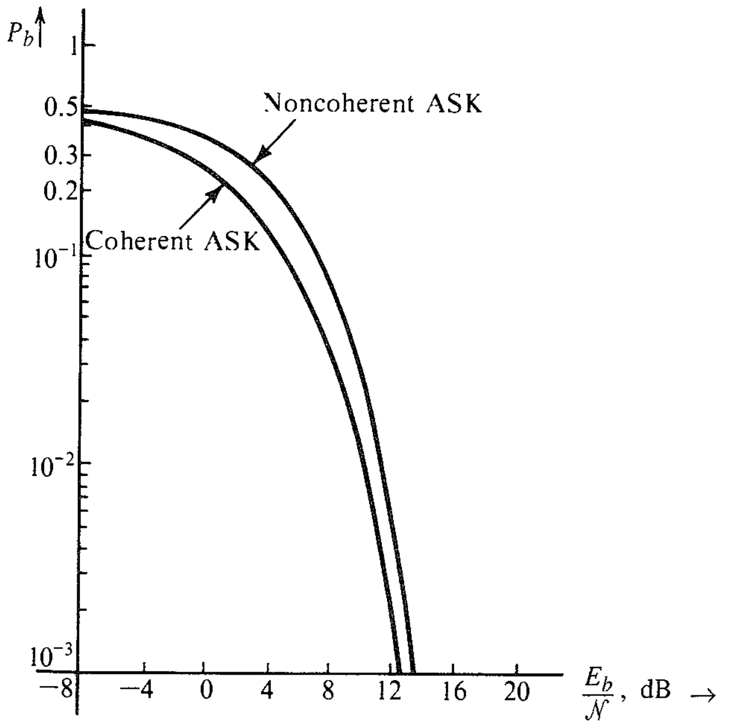


Figure 13.13 Error probability of ASK.

Frequency-Shift Keying

The noncoherent receiver for FSK is shown in Fig. 13.14. The filters $H_0(\omega)$ and $H_1(\omega)$ are matched to the two RF pulses corresponding to **0** and **1**, respectively. The outputs of the envelope detectors at $t = T_b$ are r_0 and r_1, respectively. The noise components of outputs of filters $H_0(\omega)$ and $H_1(\omega)$ are the gaussian RVs n_0 and n_1, respectively, with $\sigma_{\mathrm{n}_0} = \sigma_{\mathrm{n}_1} = \sigma_{\mathrm{n}}$.

If **1** is transmitted ($\mathrm{m} = \mathbf{1}$), then at the sampling instant, the envelope r_1 has the rician PDF*

$$p_{\mathrm{r}_1}(r_1) = \frac{r_1}{\sigma_{\mathrm{n}}^2} e^{-(r_1^2+A_p^2)/2\sigma_{\mathrm{n}}^2} I_0\left(\frac{r_1 A_p}{\sigma_{\mathrm{n}}^2}\right)$$

and r_0 is the noise envelope with Rayleigh density

$$P_{\mathrm{r}_0}(r_0) = \frac{r_0}{\sigma_{\mathrm{n}}^2} e^{-r_0^2/2\sigma_{\mathrm{n}}^2}$$

The decision is $\mathrm{m} = \mathbf{1}$ if $\mathrm{r}_1 > \mathrm{r}_0$ and $\mathrm{m} = \mathbf{0}$ if $\mathrm{r}_1 < \mathrm{r}_0$. Hence, when binary **1** is transmitted, an error is made if $\mathrm{r}_0 > \mathrm{r}_1$,

$$P(\epsilon|\mathrm{m} = \mathbf{1}) = P(\mathrm{r}_0 > \mathrm{r}_1)$$

The event $\mathrm{r}_0 > \mathrm{r}_1$ is the same as the joint event "r_1 has any positive value† and r_0 has a value greater than r_1." This is simply the joint event $(0 < \mathrm{r}_1 < \infty,\ \mathrm{r}_0 > \mathrm{r}_1)$. Hence,

$$\begin{aligned} P(\epsilon|\mathrm{m} = \mathbf{1}) &= P(0 < \mathrm{r}_1 < \infty,\ \mathrm{r}_0 > r_1) \\ &= \int_0^\infty \int_{r_1}^\infty p_{\mathrm{r}_1\mathrm{r}_0}(r_1, r_0)\, dr_1\, dr_0 \end{aligned}$$

Because r_1 and r_0 are independent, $p_{\mathrm{r}_1\mathrm{r}_0} = p_{\mathrm{r}_1} p_{\mathrm{r}_0}$. Hence,

$$\begin{aligned} P(\epsilon|\mathrm{m} = \mathbf{1}) &= \int_0^\infty \frac{r_1}{\sigma_{\mathrm{n}}^2} e^{-(r_1^2+A_p^2)/2\sigma_{\mathrm{n}}^2} I_0\left(\frac{r_1 A_p}{\sigma_{\mathrm{n}}^2}\right) \int_{\mathrm{r}_1}^\infty \frac{r_0}{\sigma_{\mathrm{n}}^2} e^{-r_0^2/2\sigma_{\mathrm{n}}^2}\, dr_1\, dr_0 \\ &= \int_0^\infty \frac{r_1}{\sigma_{\mathrm{n}}^2} e^{-(2r_1^2+A_p^2)/2\sigma_{\mathrm{n}}^2} I_0\left(\frac{r_1 A_p}{\sigma_{\mathrm{n}}^2}\right) dr_1 \end{aligned}$$

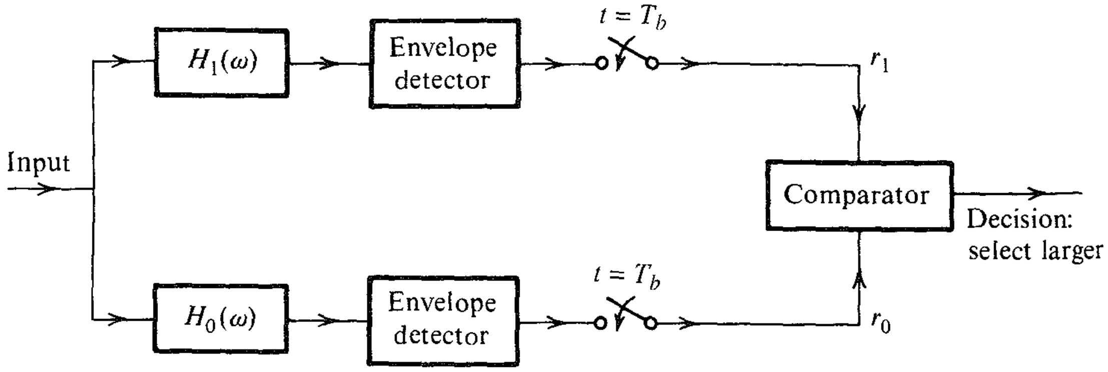


Figure 13.14 Noncoherent detection of binary FSK.

---

* An orthogonal FSK is assumed. This ensures that r_0 and r_1 have Rayleigh and Rice densities, respectively, when **1** is transmitted.

† r_1 is the envelope detector and can take only positive values.

Letting $x = \sqrt{2}\, r_1$ and $\alpha = A_p/\sqrt{2}$, we have

$$P(\epsilon|\mathrm{m} = \mathbf{1}) = \tfrac{1}{2} e^{-A_p^2/4\sigma_\mathrm{n}^2} \int_0^\infty \frac{x}{\sigma_\mathrm{n}^2} e^{-(x^2+\alpha^2)/2\sigma_\mathrm{n}^2} I_0\left(\frac{x\alpha}{\sigma_\mathrm{n}^2}\right) dx$$

Observe that the integrand is a rician density, and, hence, its integral is unity. Therefore,

$$P(\epsilon|\mathrm{m} = \mathbf{1}) = \tfrac{1}{2}\, e^{-A_p^2/4\sigma_\mathrm{n}^2} \tag{13.34a}$$

Note that for a matched filter,

$$\rho_\mathrm{max}^2 = \frac{A_p^2}{\sigma_\mathrm{n}^2} = \frac{2E_p}{\mathcal{N}}$$

For FSK, $E_b = E_p$, and Eq. (13.34a) becomes

$$P(\epsilon|\mathrm{m} = \mathbf{1}) = \tfrac{1}{2}\, e^{-\frac{1}{2}E_b/\mathcal{N}} \tag{13.34b}$$

Similarly,

$$P(\epsilon|\mathrm{m} = \mathbf{0}) = \tfrac{1}{2}\, e^{-\frac{1}{2}E_b/\mathcal{N}} \tag{13.34c}$$

and

$$P_b = \tfrac{1}{2}\, e^{-\frac{1}{2}E_b/\mathcal{N}} \tag{13.35}$$

This behavior is similar to that of noncoherent ASK [Eq. (13.32c)]. Again we observe that for $E_b/\mathcal{N} \gg 1$, the performance of coherent and noncoherent FSK are essentially similar.

From the practical point of view, FSK is to be preferred over ASK because FSK has a fixed optimum threshold, whereas the optimum threshold of ASK depends on $E_b/\mathcal{N}$ (the signal level). Hence, ASK is particularly susceptible to signal fading. Because the decision requires a comparison between r_0 and r_1, this problem does not arise in FSK. This is the outstanding advantage of noncoherent FSK over noncoherent ASK. In addition, unlike noncoherent ASK, probabilities $P(\epsilon|\mathrm{m} = \mathbf{1})$ and $P(\epsilon|\mathrm{m} = \mathbf{0})$ are equal in noncoherent FSK. The disadvantage of FSK is that it requires a larger bandwidth than ASK.

Differentially Coherent PSK

Just as it is impossible to demodulate a DSB-SC signal with an envelope detector, it is also impossible to demodulate PSK (which is really DSB-SC) noncoherently. We can, however, demodulate PSK without the synchronous, or coherent, local carrier by using what is known as **differentially coherent PSK (DPSK)**.

The optimum receiver is shown in Fig. 13.15. This receiver is very much like a correlation detector (Fig. 13.3), which is equivalent to a matched-filter detector. In a correlation detector, we multiply pulse $p(t)$ by a locally generated pulse $p(t)$. In the case of DPSK, we take advantage of the fact that the two RF pulses used in transmission are identical except for the sign. In the detector in Fig. 13.15, we multiply the incoming pulse by the preceding pulse. Hence, the preceding pulse serves as a substitute for the locally generated pulse. The only difference is that the preceding pulse is noisy because of channel noise, and this tends to degrade the performance in comparison to coherent PSK. When the output r is positive, the present pulse is identical to the previous one, and when r is negative, the present pulse is the negative of the previous pulse. Hence, from the knowledge of the first reference digit, it is possible to detect all the received digits. Detection is facilitated by using so-called *differential coding*, as discussed in Sec. 7.3.

In order to derive the DPSK error probability, we observe that DPSK using differential coding is essentially an orthogonal signaling scheme. A binary **1** is transmitted by a sequence of two pulses $(p,\ p)$ or $(-p,\ -p)$ over $2T_b$ seconds (no transition). Similarly, a binary **0** is transmitted by a sequence of two pulses $(p,\ -p)$ or $(-p,\ p)$ over $2T_b$ seconds (transition). Either of the pulse sequences used for binary **1** is orthogonal to either of the pulse sequences used for binary **0**. Because no local carrier is generated for demodulation, the detection is noncoherent, with an effective pulse energy equal to $2E_p$ (twice the energy of pulse p). The actual energy transmitted per digit is only E_p, however, the same as in noncoherent FSK. Consequently, the performance of DPSK is 3 dB superior to that of noncoherent FSK. Hence from Eq. (13.35), we can write P_b for DPSK as

$$P_b = \frac{1}{2}e^{-E_b/\mathcal{N}} \tag{13.36}$$

This error probability (Fig. 13.16) is superior to that of noncoherent FSK by 3 dB and is essentially similar to coherent PSK for $E_b/\mathcal{N} \gg 1$ [Eq. (13.25b)]. This is as expected, because we saw earlier that DPSK appears similar to PSK. Rigorous derivation of Eq. (13.36) can be found in the literature.[6]

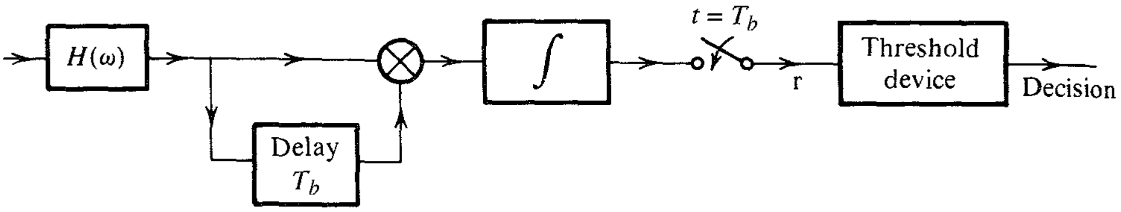


Figure 13.15 Differential PSK detection.

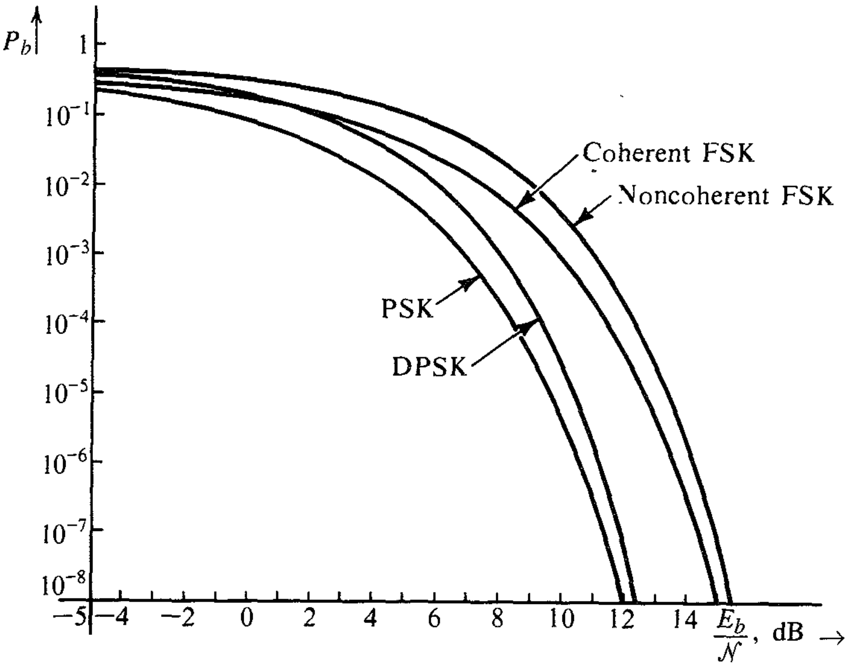


Figure 13.16 Error probability of PSK, DPSK, and coherent and noncoherent FSK.

13.4 PERFORMANCE OF SPREAD SPECTRUM SYSTEMS

In Sec. 9.2, we explained how spread spectrum systems work. Spread spectrum is basically a broad-band modulation scheme where the baseband spectrum is spread by using a pseudorandom (PN) code. At the receiver, a replica of the same PN code is used to "despread" the received signal to recover the original data. The key concept in spread spectrum systems is the PN code sequence, which is generally periodic and consists of periodic coded sequence of **1**'s and **0**'s with interesting autocorrelation property.

13.4.1 PN Sequence Generation

A good PN sequence is characterized by an autocorrelation that is similar to that of a white noise. This means the autocorrelation function of a PN sequence should be high near $\tau = 0$ and low for all $\tau \neq 0$, as shown in Fig. 13.17a. Moreover, in CDMA applications several users share the same band using different PN sequences. Hence, it is necessary that the crosscorrelation among different pairs of PN sequences be small to reduce mutual interference.

A PN code is periodic. A shift-register network with output feedback can generate a sequence with long periods and low susceptibility to structural identification by an outsider. The most widely known binary PN sequences are the **maximum-length** shift-register sequences (m-sequences). Such a sequence can be generated by an m-stage shift register with suitable feedback connections and has a length $n = 2^m - 1$ bits, which is a the maximum period for such a finite state machine. Figure 13.17b shows a shift-register encoder for $m = 6$ and $n = 63$.

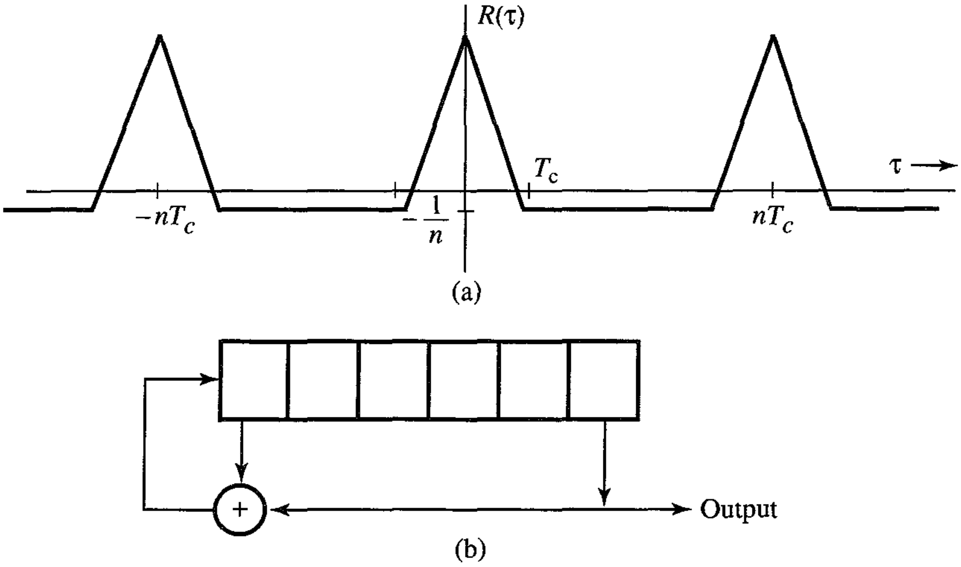


Figure 13.17 (a) PN sequence autocorrelation function. (b) Six-stage generator of a maximum-length PN sequence.

13.4.2 Optimum Detection of DS/SS PSK

The signal m(t) is a polar signal consisting of a sequence of NRZ pulses of duration T_b and constant amplitude 1 or −1.* This signal is multiplied by the PN code $c(t)$ and then DSB-modulated with a carrier $\sqrt{2}\cos(\omega_c t + \theta)$.† Although the message signal here is m(t) with binary values ±1, it is helpful to view the situation in a different way. During each bit interval T_b, we are transmitting either $\sqrt{2}c(t)\cos(\omega_c t + \theta)$ or $-\sqrt{2}c(t)\cos(\omega_c t + \theta)$. Hence, the basic binary pulse may be taken as $\sqrt{2}c(t)\cos(\omega_c t + \theta)$ [instead of m(t)], and the optimum correlation receiver would be of the form shown in Fig. 13.18 (see Fig. 13.3). The integrator integrates over the interval 0 to T_b, and its output is sampled at $t = T_b$ to decide whether the transmitted pulse was $\sqrt{2}c(t)\cos(\omega_c t + \theta)$ or $-\sqrt{2}c(t)\cos(\omega_c t + \theta)$; that is, whether m($t$) was 1 or −1. The low-pass filter shown in the receiver is not necessary because the subsequent integrator itself acts as a low-pass filter. We have shown the filter explicitly to simplify understanding of the receiver.

In Eq. (13.21b) we showed that for a channel with (white) noise of PSD $\mathcal{N}/2$, the polar optimum receiver error probability is given by

$$P_b = Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right) \tag{13.37}$$

where E_b is the energy per bit (energy of one pulse). This result appears rather surprising in view of the fact that the error probability of an optimum receiver is unchanged whether we use spread spectrum or not. We have seen in Sec. 9.2 how a spread spectrum system can reduce the power of an interfering (jamming) signal by N, the processing gain. Is noise different from an intentional interference? A heuristic explanation of this apparent anomaly lies in the fact that the spread spectrum reduces the interfering signal power by spreading its spectrum

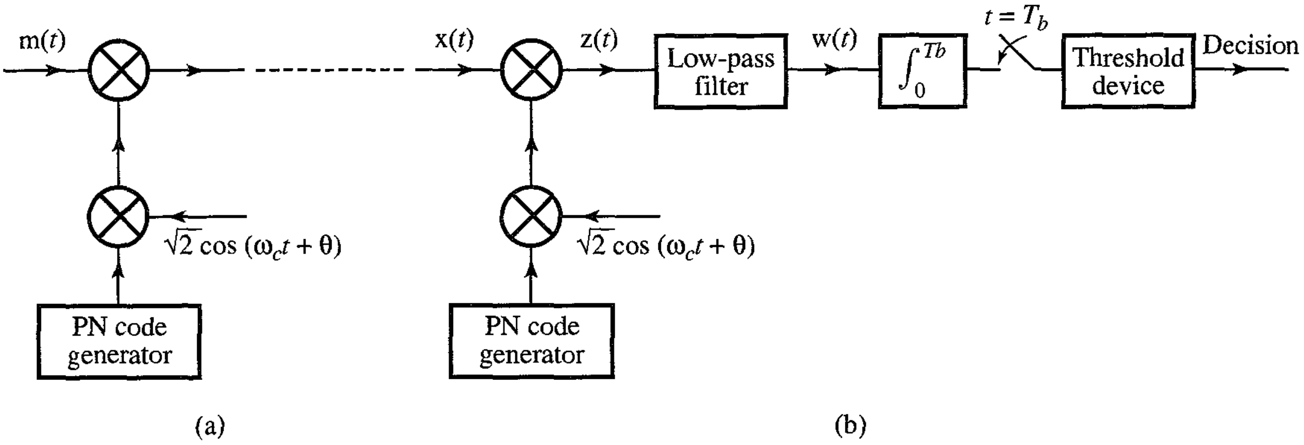


Figure 13.18 DS/SS PSK system. (a) Transmitter. (b) Receiver.

---

* In practice, the amplitude may be ±A. This makes no difference for the optimum receiver.

† The multipliers $\sqrt{2}$ in the carriers of the transmitter and the receiver in Fig. 13.18 are for convenience only. The final results (for error probability) are independent of the values of these multipliers. The choice of $\sqrt{2}$ is dictated by the fact that it leaves the power of the modulated (or the demodulated) signal identical to the signal before modulation (or demodulation). In the literature, the multiplier at the transmitter is often chosen as $\sqrt{2P}$ to ensure the transmitted signal power to be P.

over a very wide band. But the noise is a broad-band signal to begin with, and it cannot be effectively spread any more. The spectral spreading can be realized only if the interference signal bandwidth is on the order of the baseband signal.

In Sec. 9.2 we showed that a spread spectrum system reduces the power of the interfering signal by $N = T_b/T_c$. Hence, the effective noise power is the channel noise power plus the jamming signal power divided by N.

13.4.3 Multiple-Access Performance of DS/SS*

To analyze a DS/SS system with K multiple-access users, we compute the interference at the output of a certain receiver caused by the remaining $K - 1$ users. Let $\mathrm{m}_1(t)$ and $\mathrm{m}_k(t)$ be the data signals and $c_1(t)$ and $c_k(t)$ the spreading signals (PN codes) for the first user and the kth user, respectively. Let their respective carriers be $\sqrt{2}\cos(\omega_c t + \theta_1)$ and $\sqrt{2}\cos(\omega_c t + \theta_k)$. Moreover the two received signals will have different delays. Let the relative delay of the kth signal with respect to the first signal be τ_k. Hence, the signal $\mathrm{x}_k(t)$ from the kth user at the input of the receiver of the first user is

$$\begin{aligned}\mathrm{x}_k(t) &= \sqrt{2}\,\mathrm{m}_k(t-\tau_k)c_k(t-\tau_k)\cos[\omega_c(t-\tau_k)+\theta_k] \\ &= \sqrt{2}\mathrm{m}_k(t-\tau_k)c_k(t-\tau_k)\cos(\omega_c t+\phi_k) \qquad \phi_k = \theta_k - \omega_c\tau_k\end{aligned}$$

The receiver multiplies $\mathrm{x}_k(t)$ with $\sqrt{2}c_1(t)\cos(\omega_c t + \theta_1)$ to yield

$$\mathrm{z}_k(t) = 2\,\mathrm{m}_k(t-\tau_k)c_k(t-\tau_k)c_1(t)\cos(\omega_c t+\phi_k)\cos(\omega_c t+\theta_1) \tag{13.38}$$

The high-frequency component is suppressed by the low-pass filter to yield

$$\mathrm{w}_k(t) = \mathrm{m}_k(t-\tau_k)c_k(t-\tau_k)c_1(t)\cos(\phi_k-\theta_1) \tag{13.39}$$

The output (interference) of the integrator sampler is

$$I_k = \cos(\phi_k-\theta_1)\int_0^{T_b}\mathrm{m}_k(t-\tau_k)c_k(t-\tau_k)c_1(t)\,dt \tag{13.40a}$$

Recall that $\mathrm{m}_k(t)$ is a polar NRZ binary signal with amplitudes ± 1. The bit duration is T_b, and there is a possible change of amplitude of $\mathrm{m}_k(t-\tau_k)$ at $t = \tau_k$. Let this amplitude be b_{-1} before $t = \tau_k$ and b_0 after $t = \tau_k$. Then

$$I_k = \cos(\phi_k-\theta_1)\left[b_{-1}\int_0^{\tau_k}c_k(t-\tau_k)c_1(t)\,dt + b_0\int_{\tau_k}^{T_b}c_k(t-\tau_k)c_1(t)\,dt\right] \tag{13.40b}$$

This is the interference from the kth user at the output of the first user. This expression makes explicit the dependence of the interference on the crosscorrelation of the spreading codes $c_1(t)$ and $c_k(t)$. Ideally, we would like $\int c_k(t-\tau_k)c_1(t)\,dt = 0$ for any value of τ_k. This is not possible in practice, but it is possible to find many sets of sequences with good crosscorrelation properties.[7]

The Gaussian Approximation

The **multiple-access interference** (**MAI**) can be modeled as the sum of independent Bernoulli trials. One frequent approximation is to apply the central limit theorem, which implies that the sum of these tiny effects (interferences) tends toward a gaussian distribution. Since there are $K-1$ independent identically distributed interferers, the total MAI, $I=\sum_{k=2}^{K} I_k$, may be approximated by a gaussian random variable. It can be shown that the resulting bit error rate P_b is given by[8,9]

---

* Rest of the Section 13.4 is based on the material contributed by Prof. B. D. Woerner and R. M. Buehrer.

$$P_b = Q\left(\frac{1}{\sqrt{(K-1)/3N + \mathcal{N}/2E_b}}\right) \tag{13.41}$$

where N is the processing gain. We can derive this result heuristically as follows. The power of the interferer's signal $\mathrm{m}_k(t)$ is E_bT_b. But because this signal is despread by the first receiver, the power of the interfering signal $\mathrm{w}_k(t)$ is only E_bT_b/N at the first receiver. Moreover, the bandwidth of a polar NRZ binary signal is T_b Hz. Hence, the PSD of the interfering signal $\mathrm{w}_k(t)$ is E_b/N. Also there are $K-1$ identically distributed interfering signals at the first receiver whose PSDs would add to yield the PSD $(K-1)E_b/N$. However, each interfering signal has a different relative time delay τ_k, which effectively results in power reduction by a factor of 3 [recall the factor cos $(\phi_k-\theta_1)$ in Eqs. (13.40)] to yield the interference PSD as $(K-1)E_b/3N$. Consequently, the noise PSD at the first receiver is not $\mathcal{N}/2$, but $(K-1)E_b/3N+\mathcal{N}/2$. When we substitute this PSD in place of $\mathcal{N}/2$ in Eq. (13.37), we obtain Eq. (13.41).

Observe that when a single user is present ($K=1$), Eq. (13.41) reduces to Eq. (13.37), as expected. When the signal-to-noise ratio is very high ($E_b/\mathcal{N}\rightarrow\infty$), we obtain

$$\lim_{E_b/\mathcal{N}\rightarrow\infty} P_b = Q\left(\sqrt{\frac{3N}{K-1}}\right)$$

This shows the presence of an irreducible error floor for the MAI limited case.*

13.4.4 The Near/Far Problem

Equation (13.41) has been derived under the assumption that signals from all users are received with the same signal power. Within a mobile system, such an assumption may be reasonable for the forward link in which all transmissions from a centrally located base station originate with the same signal power and follow the same transmission path. However, this assumption is not realistic on the reverse link from the mobile to the base station because signals originate from diverse and moving locations. This effect can be partially compensated for by power control, which takes two forms. Under open loop power control, a mobile station adjusts its power based on the strength of the signal it receives from the base station. This presumes that a reciprocal relationship exists between forward and reverse links, an assumption that may not hold if the links operate in different frequency bands. As a result, closed loop power control is often required in which the base station orders changes in the mobile station's transmitted power.

---

* The gaussian approximation has limitations when used to predict system performance. While the central limit theorem implies that I will tend toward a gaussian distribution near the center of its distribution, convergence may be slow at its tails. Unfortunately, it is the extreme values of I that tend to produce errors. As a result, the gaussian distribution is optimistic for low BER.[10] The problem is worsened in situations where there is a single strong interferer or a single strong multipath component. This is of concern because low BERs are difficult to simulate, so we particularly desire analytical techniques that are accurate at low BERs. Several improvements to the gaussian approximation have been proposed which represent MAI more accurately under a wide range of conditions.[10–12]

The effects of the near-far problem and power control may be investigated using the analytic techniques described earlier with nonequal signal powers. Although the accuracy of the gaussian approximation is reduced in the case of the near-far problem, quantitative results may still be obtained because the central limit theorem may be applied to the chips from each interferer as well as the number of interferers K.

By extending the qualitative interpretation of Eq. (13.41), the effects of the near-far problem can be found as follows. If users have unequal powers, each will contribute $E_b^{(k)}/3N$ to the noise level, where $E_b^{(k)}$ is the energy per bit for user k. Adding these as before to the noise PSD, we can find a simple equation for the error probability of user 1, similar to Eq. (13.41):

$$P_b = Q\left[\frac{1}{\sqrt{\sum_{k=2}^{K} E_b^{(k)}/3E_b^{(1)}N + \mathcal{N}/2E_b^{(1)}}}\right] \tag{13.42}$$

Equation (13.42) illustrates the deleterious effect of the near-far problem. If just a single user has a significantly higher power level than the desired user, that interferer will dominate the performance. Substituting $E_b^{(k)} = E_b^{(1)}$ results in Eq. (13.41), as expected.

As an example, the effect of imperfect power control is investigated by Woerner and Cameron.[13] Field trial results have indicated that even after the operation of open and closed loop power control, the received signal level in the IS-95 digital cellular CDMA system reverse channel varies according to a log-normal distribution with a variance of 1 to 2 dB. The BER for this model is calculated using the method of Lehnert.[10] For a wireless voice transmission system, BERs of 10^{-2} to 10^{-3} are required for satisfactory performance. Thus the capacity may be considered to be the maximum number of users for which the BER does not exceed a fixed threshold. The results of Fig. 13.19 indicate that even after power control is applied, the variance of received power levels can result in capacity losses of 10 to 30% in a CDMA system.

13.4.5 Multipath Propagation

Thus far we have assumed that each user's signal is transmitted via a single path. However, the wireless channel is characterized by multipath propagation as signals reflect and scatter off objects in the transmission path. The constructive and destructive interference between arriving multipath components can result in severe fading in narrow-band systems. The ability of spread spectrum signals to reject and even exploit multipath propagation is a key advantage in the wireless environment. A spread spectrum signal can resolve multipath components if the chip duration is of the same order as the multipath delay. In narrow-band systems, delays are normally much smaller than a bit period, making them indistinguishable.

Since the desired user's spreading code will have low autocorrelation, we can treat these multipath components as additional interferers provided that they are at least one chip apart. Practical experience indicates that at low microwave frequencies, 1 MHz of spread bandwidth is required to resolve a significant amount of the multipath components, and 10 MHz of bandwidth will resolve nearly all.[14] Returning to the case of equal power levels, we can model each individual path as an interferer contributing $E_b/3N$ to the noise level. Thus, there are $K-1$ users with L paths along with $L-1$ interfering versions of the desired signal contributing

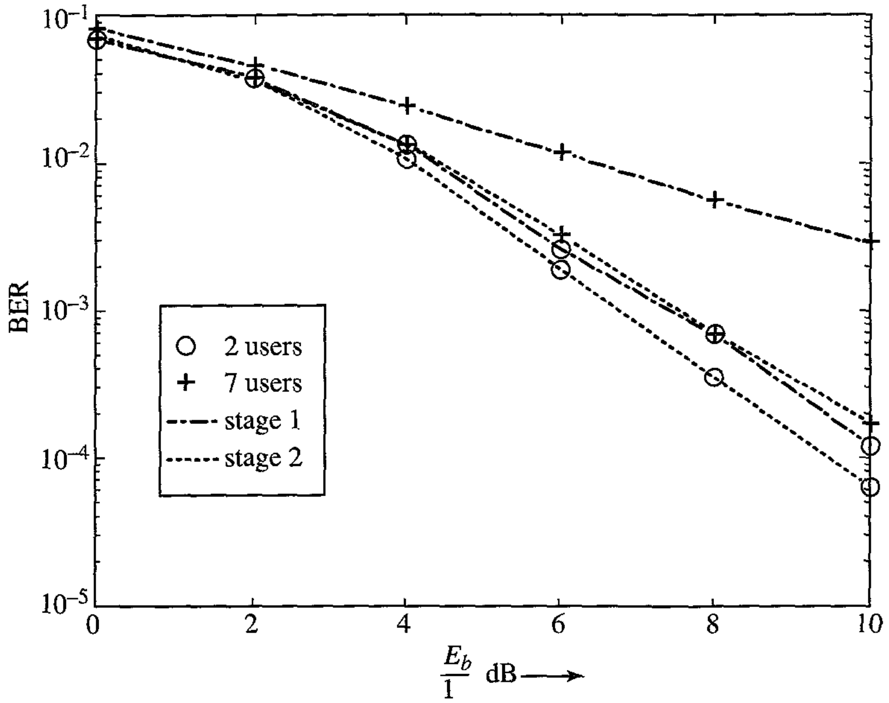


Figure 13.19 Error probability for number of users under imperfect power control.

to the noise level. This results in a total noise level of $(KL - 1)E_b/3N$ being added to the overall noise PSD. Utilizing this in Eq. (13.41) results in

$$P_b = Q\left[\frac{1}{\sqrt{(KL-1)/3N + \mathcal{N}/2E_b}}\right] \tag{13.43}$$

Though the numerical accuracy of this equation is limited, it qualitatively shows how the performance degrades as the number of arriving multipath components L increases. Also we can see that for $L = 1$, Eq. (13.43) reduces to Eq. (13.41). If there is only a single user ($K = 1$), this equation shows that performance is still below optimum due to the multipath. Equation (13.43) assumes that the multipath and interferers have power levels of the same magnitude as the desired user's primary path. Although this is rarely the case, Eq. (13.43) can be extended similarly to Eq. (13.42) for unequal power levels.

In addition to mitigating the multipath propagation, a spread spectrum may actually exploit this effect by using a Rake receiver.[15,16] A Rake receiver is a device for coherently combining the energy from two or more received multipath components, thereby increasing the received signal power and providing a form of diversity reception. The Rake consists of a bank of correlation receivers, with each individual receiver correlating with a different arriving multipath component. By adjusting for the delays, the individual multipath components can contribute correlated energy to the decision statistic, rather than harmful interference. Analogous to a garden rake, the Rake receiver gathers as much signal energy as possible.

13.4.6 Performance of Frequency Hopped Systems

Although DS/SS is currently more commercially important, some commercial applications for FH/SS are emerging. FH/SS is used in some spread spectrum devices which provide low rate data communication in the unlicensed ISM frequency band from 902 to 928 MHz, and both the cellular digital packet data (CDPD) and GSM standards have options for incorporating frequency hopping.

Normally noncoherent detection is used for FH/SS because of the difficulty in maintaining an accurate carrier phase reference under changing frequency conditions. For noncoherent detection, we have shown that [see Eq. (13.32c)]

$$P_b = \frac{1}{2} e^{-E_b/2\mathcal{N}} \tag{13.44}$$

However, if two users transmit simultaneously in the same frequency band, a collision or "hit" occurs. In this case we will assume that the probability of error is $\frac{1}{2}$. (This is actually pessimistic since studies have shown that this value can be lower.) Thus the overall probability of bit error can be modeled as

$$P_b = \frac{1}{2} e^{-E_b/2\mathcal{N}}(1 - P_h) + \frac{1}{2} P_h \tag{13.45}$$

where P_h is the probability of a hit, which we must determine. If there are J frequency slots, there is a $1/J$ probability that a given interferer will be present in the desired user's slot. If there are $K - 1$ interferers or other users, the probability that at least one is present in the desired frequency slot is

$$P_h = 1 - \left(1 - \frac{1}{J}\right)^{K-1} \approx \frac{K-1}{J} \tag{13.46}$$

assuming J is large. Substituting this into Eq. (13.45) gives

$$P_b = \frac{1}{2} e^{-E_b/2\mathcal{N}}\left(1 - \frac{K-1}{J}\right) + \frac{1}{2}\frac{K-1}{J} \tag{13.47}$$

If $K = 1$, the probability of error reduces to Eq. (13.44), the standard probability of error for BFSK. Also, if we allow $E_b/\mathcal{N}$ to approach infinity, we see that

$$\lim_{E_b/\mathcal{N} \to \infty} P_b = \frac{1}{2}\frac{K-1}{J} \tag{13.48}$$

which illustrates the irreducible error rate due to MAI.

The previous analysis assumes that all users hop their carrier frequencies synchronously. This is called **slotted** frequency hopping. This may not be a realistic scenario for many FH/SS systems. Even when synchronization can be achieved between individual user clocks, individual paths will not arrive synchronously to each other due to the various propagation delays. A simple development for asynchronous performance can be shown following the approach of Geronoitis and Pursley,[17] which shows that the probability of a hit in the asynchronous case is

$$P_h = 1 - \left[1 - \frac{1}{J}\left(1 + \frac{1}{N_b}\right)\right]^{K-1} \tag{13.49}$$

where N_b is the number of bits per hop. Comparing Eqs. (13.49) and (13.46) we see that for the asynchronous case, the probability of a hit is increased, as expected. Using Eq. (13.49) in Eq. (13.45) we obtain the probability of error for the asynchronous case as

$$P_b = \frac{1}{2} e^{-E_b/\mathcal{N}}\left[1 - \frac{1}{J}\left(1 + \frac{1}{N_b}\right)\right]^{K-1} + \frac{1}{2}\left\{1 - \left[1 - \frac{1}{J}\left(1 + \frac{1}{N_b}\right)\right]^{K-1}\right\} \tag{13.50}$$

From the preceding analysis we can draw two inferences. First, FH/SS has an advantage over DS/SS in that it is not as susceptible to the near-far problem. Because signals are generally not utilizing the same frequency simultaneously, the relative power levels of signals are not as critical as in DS/SS. The near-far problem is not totally avoided, however, since there will be some interference caused by stronger signals bleeding into weaker signals due to imperfect filtering. The second inference drawn is that to achieve reasonable throughputs, error correction coding is required. By applying strong Reed-Solomon or other burst error correcting codes, the performance can be increased dramatically.

13.5 M-ARY COMMUNICATION

Thus far we have stressed binary communication, which happens to be perhaps the single most important mode of communication in practice today. In the binary case, only two symbols are used, whereas in the M-ary case, the total number of symbols used is M. Each M-ary symbol carries as much information as $\log_2 M$ binary digits [see Eq. (7.55)]. In compensation, we may have to increase the transmitted power (as in the multiamplitude or multiphase case) or transmission bandwidth (as in the multitone case) in order to maintain a given performance level.

Multiamplitude Signaling

In the binary case, we transmit two symbols, consisting of the pulses $p(t)$ and $-p(t)$, where $p(t)$ may be either a baseband pulse or a carrier modulated by a baseband pulse. In the multiamplitude (MASK) case, the M symbols are transmitted by M pulses $\pm p(t)$, $\pm 3p(t)$, $\pm 5p(t)$, ..., $\pm(M-1)p(t)$. Thus, to transmit R_M M-ary digits per second, we are required to transmit R_M pulses per second of the form $kp(t)$. Pulses are transmitted every T_M seconds, so that $T_M = 1/R_M$. If E_p is the energy of pulse $p(t)$, then assuming that pulses $\pm p(t)$, $\pm 3p(t)$, $\pm 5p(t)$, ..., $\pm(M-1)p(t)$ are equally likely, the average pulse energy E_{pM} is given by

$$\begin{aligned} E_{pM} &= \frac{2}{M}[E_p + 9E_p + 25E_p + \cdots + (M-1)^2 E_p] \\ &= \frac{2E_p}{M}\sum_{k=0}^{\frac{M-2}{2}}(2k+1)^2 \\ &= \frac{M^2-1}{3}E_p && (13.51a) \\ &\simeq \frac{M^2}{3}E_p \qquad M \gg 1 && (13.51b) \end{aligned}$$

Recall that an M-ary symbol carries an information of $\log_2 M$ bits. Hence, the bit energy E_b is

$$E_b = \frac{E_{pM}}{\log_2 M} = \frac{M^2-1}{3\log_2 M}E_p \tag{13.51c}$$

Because the transmission bandwidth is independent of the pulse amplitude, the M-ary bandwidth is the same as in the binary case for the given rate of pulses, yet it carries more information.

This means that for a given information rate, the MASK bandwidth is less than that of the binary case by a factor of $\log_2 M$.

To calculate the error probability, we observe that because we are dealing with the same basic pulse $p(t)$, the optimum M-ary receiver is a filter matched to $p(t)$. When the input pulse is $kp(t)$, the output $\mathrm{r}(T_p)$ at the sampling instant will be $kA_p + \mathrm{n}_o(T_M)$. Note that $A_p = E_p$, the energy of $p(t)$, and that σ_n^2, the variance of $\mathrm{n}_o(t)$, is $\mathcal{N}E_p/2$. Thus, the optimum receiver for the multiamplitude M-ary signaling case is identical to that of the polar binary case (see Fig. 13.3 or 13.6a). The sampler has M possible outputs $\pm kA_p + \mathrm{n}_o(T_M)$ $(k = 1, 3, 5, \ldots, M-1)$ that we wish to detect. The conditional PDFs $p(r|m)$ are gaussian with mean kA_p and variance σ_n^2, as shown in Fig. 13.20a. Let P_{eM} be the error probability detecting a symbol and $P(\epsilon|m)$ be the error probability given that the symbol m is transmitted.

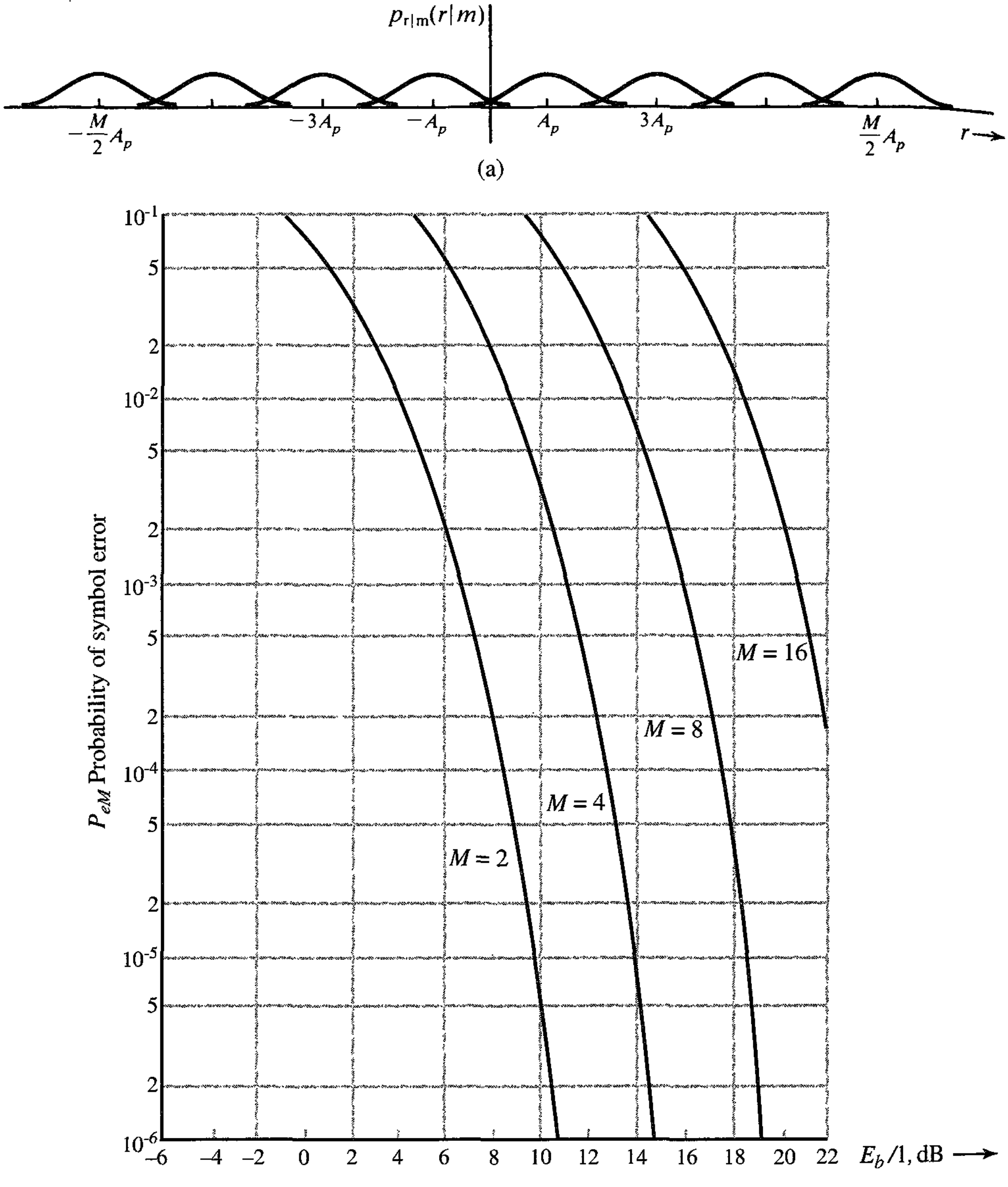
To calculate P_{eM}, we observe that the case of the two extreme symbols [represented by $\pm(M-1)p(t)$] is similar to the binary case because they have to guard against only one neighbor. As for the remaining symbols, they must guard against neighbors on both sides, and, hence, $P(\epsilon|m)$ in this case is twice that of the extreme symbol. From Fig. 13.20a it is evident that $P(\epsilon|m_i)$ is $Q(A_p/\sigma_\mathrm{n})$ for the two extreme signals and is $2Q(A_p/\sigma_\mathrm{n})$ for the remaining $(M-2)$ symbols. Hence,

$$\begin{aligned}
P_{eM} &= \sum_{i=1}^{M} P(m_i)P(\epsilon|m_i) \\
&= \frac{1}{M}\sum_{i=1}^{M} P(\epsilon|m_i) \\
&= \frac{1}{M}\left[Q\left(\frac{A_p}{\sigma_\mathrm{n}}\right) + Q\left(\frac{A_p}{\sigma_\mathrm{n}}\right) + (M-2)2Q\left(\frac{A_p}{\sigma_\mathrm{n}}\right)\right] \\
&= \frac{2(M-1)}{M} Q\left(\frac{A_p}{\sigma_\mathrm{n}}\right) \qquad (13.52\text{a})
\end{aligned}$$

For a matched-filter receiver, $(A_p/\sigma_\mathrm{n})^2 = 2E_p/\mathcal{N}$, and

$$\begin{aligned}
P_{eM} &= 2\left(\frac{M-1}{M}\right) Q\left(\sqrt{\frac{2E_p}{\mathcal{N}}}\right) \qquad (13.52\text{b}) \\
&= 2\left(\frac{M-1}{M}\right) Q\left[\sqrt{\frac{6\log_2 M}{M^2-1}\left(\frac{E_b}{\mathcal{N}}\right)}\right] \qquad (13.52\text{c}) \\
&\approx 2Q\left[\sqrt{\frac{6\log_2 M}{M^2}\left(\frac{E_b}{\mathcal{N}}\right)}\right] \qquad M \gg 1 \qquad (13.52\text{d})
\end{aligned}$$

Bit Error Rate (BER): It is somewhat unfair to compare M-ary signaling on the basis of P_{eM}, the error probability of an M-ary symbol, which conveys the information of $k = \log_2 M$ bits. This weighs unfairly against higher values of M. For a fair comparison, we should compare various schemes for P_b, the probability of bit error, rather than P_{eM}, the probability of symbol error (symbol error rate). We now show that for multiamplitude signaling $P_b = P_{eM}/\log_2 M$.



(a)

(b)

Figure 13.20 (a) Conditional PDFs in MASK. (b) Error probability in MASK.

Because the type of errors that predominate are those where a symbol is mistaken for its immediate neighbors (see Fig. 13.20a), it would be logical to assign neighboring M-ary symbols, binary code words that differ in the least possible digits. The Gray code* is suitable for this because adjacent binary combinations in this code differ only by one digit. Hence, an error in one M-ary symbol detection will cause only one error in a group of $\log_2 M$ binary digits transmitted by the M-ary symbol. Hence, the bit error rate $P_b = P_{eM}/\log_2 M$. However, the factor $\log_2 M$ is negligible compared to other factors that vary exponentially with M in the expression for P_{eM} [see Eqs. (13.52)], and, hence, it is not as influential in determining the behavior of P_{eM}. Figure 13.20b shows P_{eM} as a function of $E_b/\mathcal{N}$ for several values of M. The relationship $P_b = P_{eM}/\log_2 M$, although valid for MASK, is not valid for other schemes due to the specific code structure. One must recompute this relationship for each scheme.

---

* The Gray code can be constructed as follows: Construct an n-digit natural binary code (NBC) corresponding to 2^n decimal numbers. If $b_1 b_2 \ldots b_n$ is a code word in this code, then the corresponding Gray code word $g_1 g_2 \cdots g_n$ is obtained by the rule

Trade-off between Power and Bandwidth

To maintain a given information rate, the pulse transmission rate in the M-ary case is reduced by the factor $k = \log_2 M$. This means the bandwidth of the M-ary case is reduced by the same factor $k = \log_2 M$. But to maintain the same P_{eM}, Eqs. (13.52) show that the power transmitted (which is proportional to E_b) increases roughly as $M^2/\log_2 M = 2^{2k}/k$. On the other hand, if we maintain a given bandwidth, the information rate in the M-ary case is increased by the factor $k = \log_2 M$. The transmitted power is equal to E_b times the bit rate. Hence, an increased data rate also necessitates increased power by the factor $(M^2/\log_2 M)(\log_2 M) = 2^{2k}$. Thus, the power increases exponentially with the increase in information rate by a factor of k. In high-powered radio systems, such a power increase may not be tolerable. Multiamplitude systems are attractive where bandwidth is at a premium. Because the voice channels of a telephone network have a fixed bandwidth, multiamplitude (or multiphase, or a combination of both) signaling appears to be an attractive method of increasing the information rate, particularly because telephone lines can be made to have a high degree of channel stability and low additive noise.

All the results derived here apply to baseband as well as modulated digital systems with coherent detection. For noncoherent detection, similar relationships exist between the binary and M-ary systems.*

Multiphase Signaling

In binary PSK (BPSK), the two basic pulses are $p'(t)\cos\omega_c t$ and $p'(t)\cos(\omega_c t + \pi)$. We generalize the idea for the M-ary case (MPSK) using M pulses with the kth pulse $p'(t)\cos[\omega_c t + (2\pi/M)k]$. Thus, the phases of successive pulses are $2\pi/M$ radians apart (Fig. 13.21). From the symmetry of this scheme it is evident that if, because of channel noise, the phase of any pulse deviates by more than π/M radians, an error is made. Hence, the error probability P_{eM} is the probability that the phase of any pulse deviates by more than π/M radians.

At the receiver, coherent phase reference is available. The basic function of the receiver is to detect the phase of the received pulse. This can be done by two phase detectors with cos

---

$$g_1 = b_1$$
$$g_k = b_k \oplus b_{k-1} \qquad k \geq 2$$

Thus for $n = 3$, the binary code **000, 001, 010, 011, 100, 101, 110, 111** is transformed into the Gray code **000, 001, 011, 010, 110, 111, 101, 100**.

* For the noncoherent case, the baseband pulses must be of the same polarity; for example, 0, $p(t)$, $2p(t)$, ..., $(M-1)p(t)$.

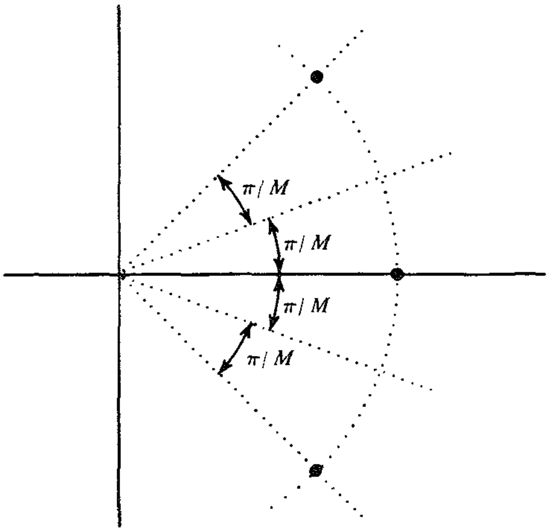


Figure 13.21 MPSK signals.

$\omega_c t$ and $\sin \omega_c t$ as respective references, along with a logic circuit to determine the ratio of the two detected components.*

To compute the error probability P_{eM}, we note that a detection error results if the phase of any pulse deviates by more than π/M. The PDF $p_\Theta(\theta)$ of the phase Θ of a sinusoid plus a bandpass gaussian noise is found in Eq. (11.64d). Hence,

$$P_{eM} = 1 - \int_{-\pi/M}^{\pi/M} p_\Theta(\theta)$$

The PDF $p_\Theta(\theta)$ in Eq. (11.64d) involves A (the sinusoid amplitude) and σ_n^2 (the noise variance). Assuming matched filtering and white noise [see Eq. (13.9c)],

$$\frac{A^2}{\sigma_n^2} = \frac{2E_p}{\mathcal{N}} = \frac{2E_b \log_2 M}{\mathcal{N}}$$

Hence,

$$P_{eM} = 1 - \frac{1}{2\pi} \int_{-\pi/M}^{\pi/M} e^{-\frac{E_b \log_2 M}{\mathcal{N}}} \left\{ 1 + \sqrt{\frac{4\pi E_b \log_2 M}{\mathcal{N}}} \cos\theta \, e^{\frac{E_b \cos^2\theta \log_2 M}{\mathcal{N}}} \times \left[1 - Q\left(\sqrt{\frac{2E_b \log_2 M}{\mathcal{N}}} \cos\theta \right) \right] \right\} d\theta \qquad (13.53)$$

Figure 13.22 shows the plot of P_{eM} as a function of $E_b/\mathcal{N}$. For $E_b/\mathcal{N} \gg 1$ (weak noise) and $M \gg 2$, Eq. (13.53) can be approximated as[18]

$$P_{eM} \simeq 2Q\left(\sqrt{\frac{2E_b \log_2 M}{\mathcal{N}}} \sin\frac{\pi}{M} \right) \qquad (13.54a)$$

---

[*] This can be done by a parallel bank of two filters matched to $p'(t) \sin \omega_c t$ and $p'(t) \cos \omega_c t$, respectively. If the input to this bank is $p'(t) \cos(\omega_c t + \theta)$, then the outputs of the two matched filters are proportional to $\sin\theta$ and $\cos\theta$, respectively. This is because the input $p'(t)\cos(\omega_c t + \theta)$ can be expressed as a sum of two orthogonal components $\sin\theta \, p'(t) \sin\omega_c t$ and $\cos\theta \, p'(t) \cos \omega_c t$. The ratio of the outputs of these filters is $\tan\theta$.

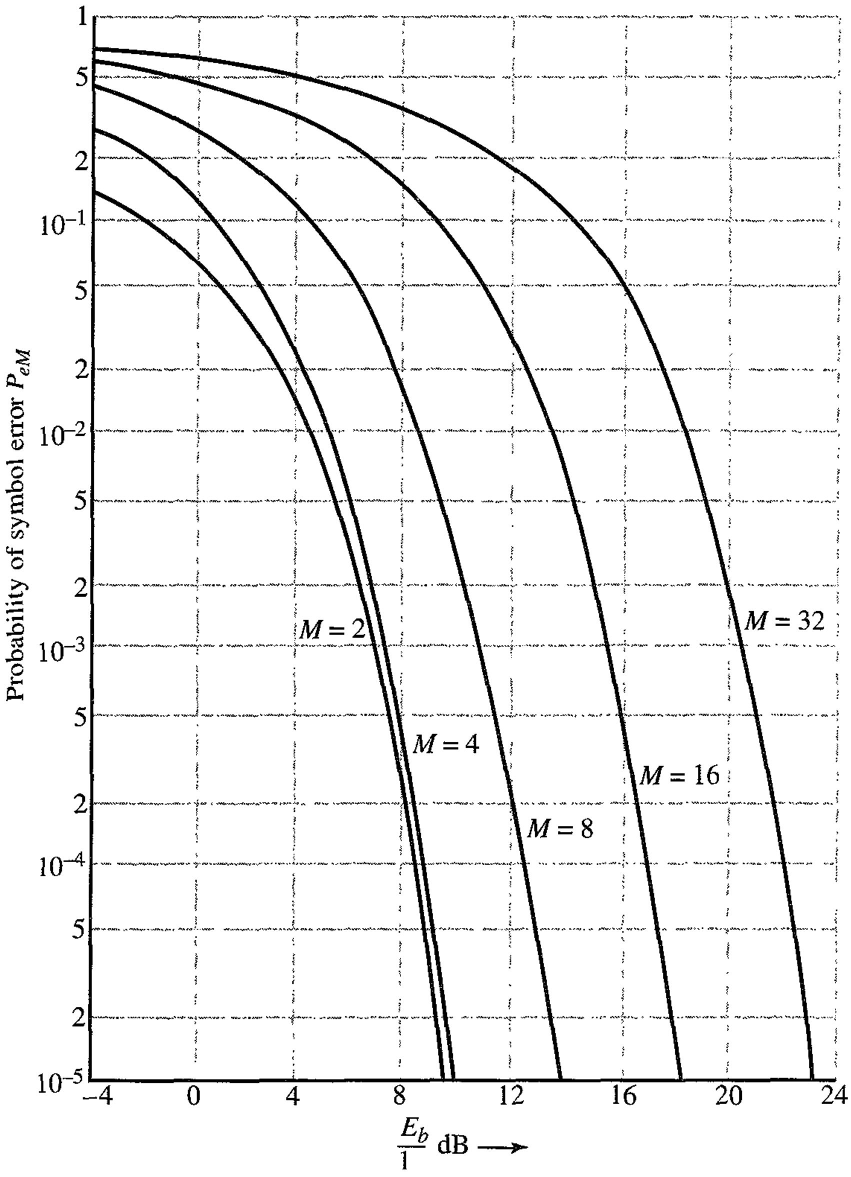


Figure 13.22 Error probability of MPSK.

$$\simeq 2Q\left(\sqrt{\frac{2\pi^2 E_b \log_2 M}{M^2 \mathcal{N}}}\right) \tag{13.54b}$$

An alternate expression for P_{eM} is derived in Chapter 14 [Eq. (14.59b)]. A comparison of Eq. (13.54b) with Eq. (13.52d) shows that the behavior of MPSK is very similar to that of MASK, at least for large $E_b/\mathcal{N}$ and $M \gg 2$. Hence, the comments made in reference to MASK apply to MPSK also.

One of the schemes frequently used in practice is the case of $M = 4$. This is four-phase, or quadriphase, PSK (QPSK). This case can be thought of as two binary PSK systems in parallel in which the carriers are in phase quadrature. Thus, one system uses pulses $\pm p(t)\cos\omega_c t$, and the other uses pulses $\pm p(t)\sin\omega_c t$. Because the carriers are in quadrature, both of these signals can be transmitted over the same channel without interference. Addition of these two streams yields four possible pulses, $\pm\sqrt{2}\,p(t)\cos(\omega_c t \pm \pi/4)$. This is precisely QPSK. Thus, QPSK doubles the transmission rate without increasing the bandwidth.

Bit Error Rate (BER): As in the case of MASK, the BER for MPSK also is given by $P_b = P_{eM}/\log_2 M$. This is because in MASK and MPSK, the type of errors that predominate are those where a symbol is mistaken for its immediate neighbor. Using Gray code we can assign the adjacent symbols codes that differ in just one digit.

An example of QPSK is found in the SPADE multiple-access communication system used for PCM via the Intelsat satellite communication global network, which attains a rate of 64 kbit/s over a transmission bandwidth of 38 kHz. The scheme uses Nyquist's criterion with a roll-off factor of 19% (see Example 7.1). Regular telephone lines with a bandwidth on the order of 2400 Hz (from 600 to 3000 Hz) are used to transmit binary data at a rate of 1200 digits per second. This rate can be increased by using MPSK. For example, to transmit data at respective rates of 2400 and 4800 digits per second, QPSK and 8-ary PSK are used. Higher rates (9600 digits per second) are obtained by using 16-ary QAM (discussed next).

Quadrature Amplitude Modulation

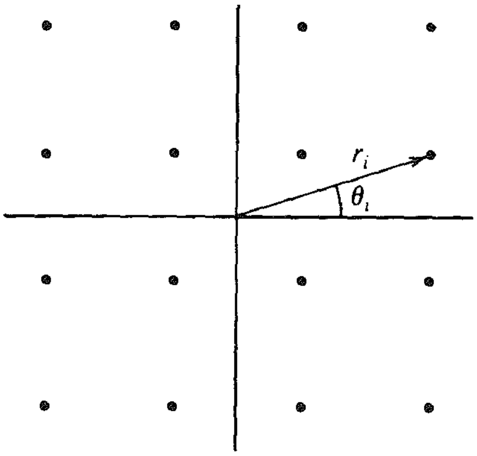
The **quadrature amplitude modulation (QAM)**, also called **amplitude phase-shift keying (APK)**, is a combination of MASK and MPSK. A QAM signal can be written as

$$\begin{aligned} p_i(t) &= p'(t)(a_i \cos \omega_c t + b_i \sin \omega_c t) \\ &= p'(t)[r_i \cos(\omega_c t + \theta_i)] \qquad i = 1, 2, \ldots, M \end{aligned} \tag{13.55}$$

where $r_i = \sqrt{a_i^2 + b_i^2}$ and $\theta_i = -\tan^{-1} b_i/a_i$. In MASK, $\theta_i = 0$ for all i; only r_i is different. In MPSK, r_i is constant for all i; only θ_i is different. In APK, both r_i and θ_i vary with i. Graphically, signal $p_i(t)$ can be mapped into a point $(a_i,\ b_i)$ or $(r_i,\ \theta_i)$. One such signal set is shown in Fig. 13.23. Because each M-ary pulse consists of a sum of pulses modulated by carriers in quadrature (quadrature multiplexing), this is the case of QAM.

Techniques for analyzing such general signals are discussed in Chapter 14. For the 16-ary case in Fig. 13.23, it can be shown that (see Example 14.3)

$$P_{eM} \simeq 3Q\left(\sqrt{\frac{4E_b}{5\mathcal{N}}}\right) \tag{13.56}$$



(b)

Figure 13.23 16-ary (or 16-point) QAM.

General analysis of MQAM (M-ary QAM) is rather involved and requires advanced techniques discussed in Chapter 14, where a 16-ary QAM is analyzed in Example 14.3. It can be shown that[18] an upper bound on P_{eM} is given by

$$P_{eM} = 4Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}\eta_M}\right)$$

where η_M is an efficiency factor, normalized to the efficiency of antipodal signaling, and depends on the constellation size and shape with $\eta_M \leq 1$. Wilson[18] gives the following values of n_M for various values of M for MASK and MQAM:

MASK	MQAM	η_MdB
4	16	−4
—	32	−6
8	64	−8.5
—	128	−10.2
16	256	−13.3

Multitone Signaling (MFSK)

In this case, M symbols are transmitted by M orthogonal pulses of frequencies $\omega_1, \omega_2, \ldots, \omega_M$, each of duration T_M. Thus, the M transmitted pulses are of the form* $\sqrt{2}p'(t)\cos\omega_k t$, where $\omega_k = 2\pi(N+k)/T_M$. The receiver (Fig. 13.24) is a simple extension of the binary receiver. The incoming pulse is multiplied by the corresponding references $\sqrt{2}\cos\omega_i t$ $(i = 1, 2, \ldots, M)$. The filter $H(\omega)$ is matched to the baseband pulse $p'(t)$. The same result is obtained if in the

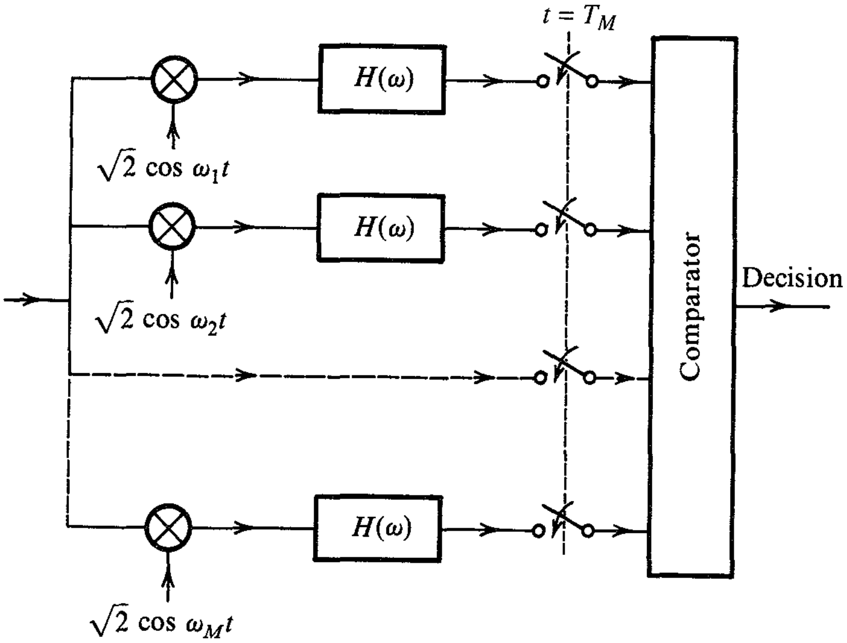


Figure 13.24 Coherent MFSK receiver.

---

* A better scheme is to use pulses $\sqrt{2}p'(t)\sin\omega_k t$ and $\sqrt{2}p'(t)\cos\omega_k t$, $k = 1, 2, \ldots, M/2$. Because these are all orthogonal pulses, we required only $M/2$ distinct frequencies, and the bandwidth is reduced by a factor of 2. This scheme cannot be used for noncoherent detection for obvious reasons.

ith bank, instead of using a multiplier and $H(\omega)$, we use a filter matched to the RF pulse $p'(t)\cos\omega_i t$. The M bank outputs sampled at $t = T_M$ are $\mathrm{r}_1,\ \mathrm{r}_2,\ \ldots,\ \mathrm{r}_M$. They are compared, and the decision is $\mathrm{m} = j$ if $\mathrm{r}_j > \mathrm{r}_i$ for all $i \neq j$. If a pulse $p'(t)\cos\omega_i t$ is received, it will cause an output only in the ith bank and will be completely suppressed in the outputs of all the remaining banks if all the M RF pulses are orthogonal.*

The bandwidth of MFSK increases with M. When $\mathrm{m} = \mathbf{1}$ is transmitted, the corresponding sampler output will be $A_p + \mathrm{n}_1$, and the other sampler outputs are $\mathrm{n}_2,\ \mathrm{n}_3,\ \ldots,\ \mathrm{n}_M$, where $\mathrm{n}_1,\ \mathrm{n}_2,\ \ldots,\ \mathrm{n}_M$ are all mutually independent gaussian variables with the same variance σ_n^2. An error is made if $\mathrm{n}_j > A_p + \mathrm{n}_1\ (j \neq 1)$. Suppose the value of $\mathrm{r}_1 = r_1$. Now the correct decision will be made if $\mathrm{n}_2,\ \mathrm{n}_3,\ \ldots,\ \mathrm{n}_M$ are all less than r_1. Because all these variables have identical PDFs (gaussian with variance σ_n), the probability of $\mathrm{n}_j < r_1$ is

$$\int_{-\infty}^{r_1} \frac{1}{\sigma_\mathrm{n}\sqrt{2\pi}} e^{-n_j^2/2\sigma_\mathrm{n}^2}\, dn_j$$

If $P(C|\mathrm{m} = \mathbf{1})$ is the probability of making a correct decision given that $\mathrm{m} = \mathbf{1}$ is transmitted, then

$$\begin{aligned}
P(C|\mathrm{m} = \mathbf{1}) &= P(\mathrm{r}_1 < \infty,\ \mathrm{n}_2 < \mathrm{r}_1,\ \mathrm{n}_3 < \mathrm{r}_1, \ldots,\ \mathrm{n}_M < \mathrm{r}_1) \\
&= \int_{-\infty}^{\infty} p_{\mathrm{r}_1}(r_1)\, dr_1 \left(\prod_{j=2}^{M} \int_{-\infty}^{r_1} \frac{1}{\sigma_\mathrm{n}\sqrt{2\pi}} e^{-n_j^2/2\sigma_\mathrm{n}^2}\, dn_j \right) \\
&= \frac{1}{(2\pi\sigma_\mathrm{n}^2)^{M/2}} \int_{-\infty}^{\infty} e^{-(r_1 - A_p)^2/2\sigma_\mathrm{n}^2}\, dr_1 \left(\int_{-\infty}^{r_1} e^{-x^2/2\sigma_\mathrm{n}^2}\, dx \right)^{M-1} \\
&= \frac{1}{\sigma_\mathrm{n}\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(r_1 - A_p)^2/2\sigma_\mathrm{n}^2} \left[1 - Q\left(\frac{r_1}{\sigma_\mathrm{n}} \right) \right]^{M-1} dr_1
\end{aligned}$$

Because of symmetry, $P(C|\mathrm{m} = \mathbf{1}) = P(C|\mathrm{m} = \mathbf{2}) = \cdots = P(C|\mathrm{m} = \mathbf{M})$. Hence, P_{CM}, the probability of a correct decision, is $P(C|\mathrm{m} = \mathbf{1})$ provided all symbols are equiprobable. Because the filters are matched [Eq. (13.9)],

$$\frac{A_p^2}{\sigma_\mathrm{n}^2} = \frac{2E_p}{\mathcal{N}} = \frac{2E_b \log_2 M}{\mathcal{N}}$$

and we can express

$$P_{eM} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(y - \sqrt{2E_b \log_2 M/\mathcal{N}}\right)^2/2} [1 - Q(y)]^{M-1}\, dy \tag{13.57}$$

---

* This can be shown as follows: The filter matched to $p(t)$ has impulse response $h(t) = p(T_M - t)$. If $q(t)$ is applied at the input of this filter, the response $r(t)$ is

$$r(t) = h(t) * q(t) = \int_{-\infty}^{\infty} p(T_M - x) q(t - x)\, dx$$

and

$$r(T_M) = \int_{-\infty}^{\infty} p(T_M - x) q(T_M - x)\, dx = \int_{-\infty}^{\infty} p(t) q(t)\, dt$$

If $p(t)$ and $q(t)$ are orthogonal, $r(T_M) = 0$.

The integral appearing on the right-hand side of Eq. (13.57) is computed and plotted in Fig. 13.25 (P_{eM} vs. $E_b/\mathcal{N}$). This plot shows an interesting behavior for the case of $M = \infty$. By properly taking the limit of P_{eM} in Eq. (13.57) as $M \to \infty$, it can be shown that[3]

$$\lim_{M\to\infty} P_{eM} = \begin{cases} 1 & E_b/\mathcal{N} < \log_e 2 \\ 0 & E_b/\mathcal{N} \geq \log_e 2 \end{cases}$$

Because the signal power $S_i = E_b R_b$, where R_b is the bit rate, it follows that for error-free communication,

$$\frac{E_b}{\mathcal{N}} \geq \log_e 2 = \frac{1}{1.44} \quad \text{or} \quad \frac{S_i}{\mathcal{N} R_b} \geq \frac{1}{1.44}$$

Hence,

$$R_b \leq 1.44 \frac{S_i}{\mathcal{N}} \text{ bit/s} \tag{13.58}$$

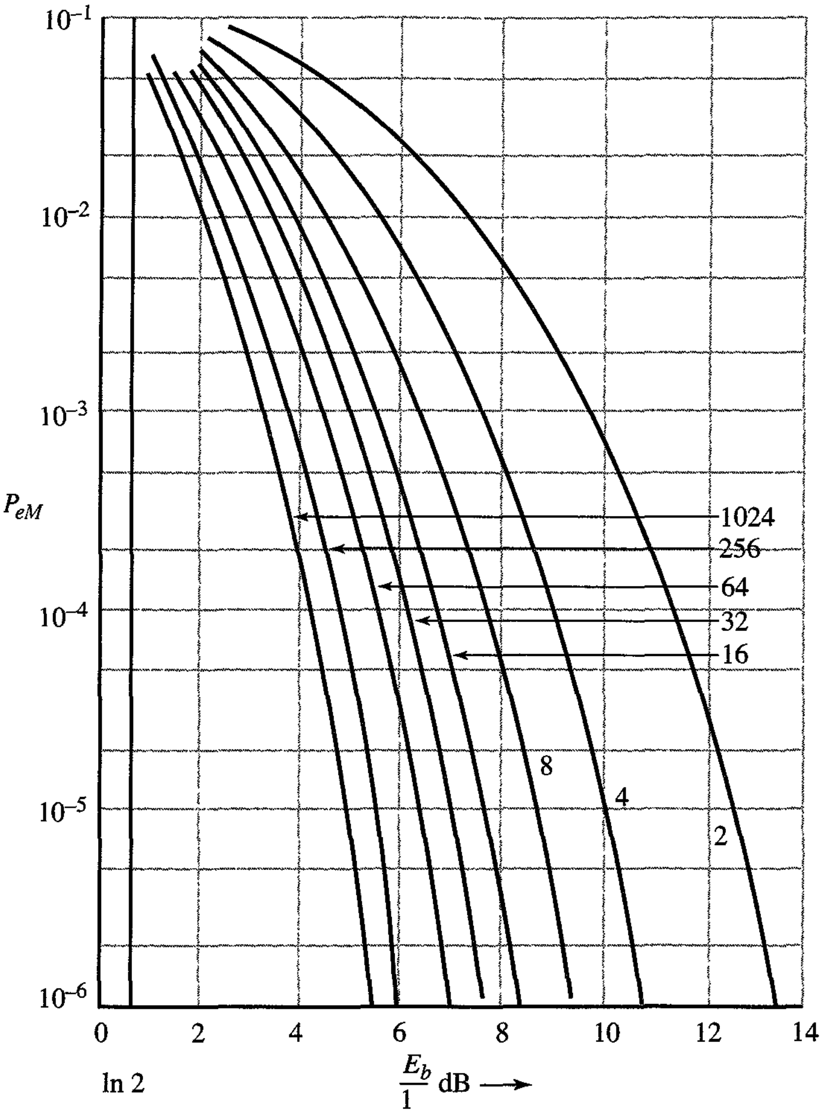


Figure 13.25 Error probability of coherent MFSK.

This shows that M-ary orthogonal signaling can transmit data error-free at a rate of up to 1.44 $S_i/\mathcal{N}$ bit/s as $M \rightarrow \infty$ (see Fig. 13.25).

Bit Error Rate (BER): For MASK and MPSK, we showed that $P_b = P_{eM}/\log_2 M$. This result is not valid for MFSK because in MASK and MPSK, the type of errors that predominate are those where a symbol is mistaken for its immediate neighbor. Using Gray code we can assign the adjacent symbols codes that differ in just one digit. In MFSK, on the other hand, a symbol is equally likely to be mistaken for any of the remaining $M - 1$ symbols. Hence, $P(\epsilon)$, the probability of mistaking one M-ary symbol for another, is

$$P(\epsilon) = \frac{P_{eM}}{M-1} = \frac{P_{eM}}{2^k - 1}$$

If an M-ary symbol differs by 1 bit from N_1 number of symbols, and differs by 2 bits from N_2 number of symbols, and so on, then $\overline{N}_\epsilon$, the average number of bits in error in reception of an M-ary symbol, is

$$\begin{aligned}
\overline{N}_\epsilon &= \sum_{n=1}^{k} n N_n P(\epsilon) \\
&= \sum_{n=1}^{k} n N_n \frac{P_{eM}}{2^k - 1} \\
&= \frac{P_{eM}}{2^k - 1} \sum_{n=1}^{k} n \binom{k}{n} \\
&= k\, 2^{k-1} \frac{P_{eM}}{2^k - 1}
\end{aligned}$$

This is an average number of bits in error in a sequence of k bits (one M-ary symbol). Consequently, the BER, P_b is this figure divided by k,

$$P_b = \frac{2^{k-1}}{2^k - 1} P_{eM} \approx \frac{P_{eM}}{2} \qquad k \gg 1$$

Noncoherent MFSK

From the practical point of view, the phase coherence of M frequencies is difficult to maintain. Hence in practice, coherent MFSK is rarely used. Noncoherent MFSK is more common. The receiver for noncoherent MFSK is similar to that for binary noncoherent FSK (Fig. 13.14), but with M banks corresponding to M frequencies. The filter $H_i(\omega)$ is matched to the RF pulse $p(t) \cos \omega_i t$. The analysis is straightforward. If $\mathrm{m} = \mathbf{1}$ is transmitted, then r_1 is the envelope of a sinusoid of amplitude A_p plus bandpass gaussian noise, and r_j ($j = 2, 3, \ldots, M$) is the envelope of the bandpass gaussian noise. Hence, r_1 has rician density, and $\mathrm{r}_2, \mathrm{r}_3, \ldots, \mathrm{r}_M$ have Rayleigh density. Using the same arguments as in the coherent case, we have

$$\begin{aligned}
P_{\mathrm{CM}} &= P(C|\mathrm{m} = \mathbf{1}) = P(0 \le \mathrm{r}_1 < \infty,\ \mathrm{n}_2 < \mathrm{r}_1,\ \mathrm{n}_3 < \mathrm{r}_1, \ldots,\ \mathrm{n}_M < \mathrm{r}_1) \\
&= \int_0^\infty \frac{\mathrm{r}_1}{\sigma_\mathrm{n}^2} I_0\left(\frac{r_1 A_p}{\sigma_\mathrm{n}^2}\right) e^{-(r_1^2 + A_p^2)/2\sigma_\mathrm{n}^2}\, dr_1 \left(\int_0^{r_1} \frac{x}{\sigma_\mathrm{n}^2} e^{-x^2/2\sigma_\mathrm{n}^2}\, dx\right)^{M-1} dx
\end{aligned}$$

$$= \int_0^\infty \frac{\mathrm{r}_1}{\sigma_\mathrm{n}^2} I_0\left(\frac{r_1 A_p}{\sigma_\mathrm{n}^2}\right) e^{-(r_1^2+A_p^2)/2\sigma_\mathrm{n}^2} \left(1 - e^{-r_1^2/2\sigma_\mathrm{n}^2}\right)^{M-1} dr_1$$

Substituting $r_1^2/2\sigma_n^2 = x$ and $(A_p/\sigma_n)^2 = 2E_p/\mathcal{N} = 2E_b \log M/\mathcal{N}$, we obtain

$$P_\mathrm{CM} = e^{-\frac{E_b \log_2 M}{\mathcal{N}}} \int_0^\infty e^{-x}(1 - e^{-x})^{M-1} I_0\left(2\sqrt{\frac{xE_b \log_2 M}{\mathcal{N}}}\right) dx \qquad (13.59a)$$

Expanding $(1 - e^x)^{M-1}$ using the binomial theorem, we obtain

$$(1 - e^x)^{M-1} = \sum_{m=0}^{M-1} \binom{M-1}{m} (-1)^m e^{-mx}$$

Substitution of this result into Eq. (13.59a) and recognizing that

$$\int_0^\infty ye^{-ay^2} I_0(by)\, dy = \frac{1}{2a} e^{-b^2/4a}$$

we obtain (after interchanging the order of summation and integration)

$$P_\mathrm{CM} = \sum_{m=0}^{M-1} \binom{M-1}{m} \frac{(-1)^m}{m+1} e^{-mE_b \log_2 M/\mathcal{N}(m+1)} \qquad (13.59b)$$

and

$$P_{eM} = 1 - P_\mathrm{CM} = \sum_{m=1}^{M-1} \binom{M-1}{m} \frac{(-1)^{m+1}}{m+1} e^{-mE_b \log_2 M/\mathcal{N}(m+1)} \qquad (13.59c)$$

The error probability P_{eM} is shown in Fig. 13.26 as a function of $E_b/\mathcal{N}$. It can be seen that the performance of noncoherent MFSK is only slightly inferior to that of coherent MFSK, particularly for large M.

Comments on M-ary Signaling

M-ary signaling provides us with additional means of exchanging, or trading, the transmission rate, transmission bandwidth, and transmitted power. It provides us a flexibility in designing a proper communication system. Thus, for a given rate of transmission, we can trade the transmission bandwidth for transmitted power. We can also increase the information rate by a factor of $k\,(k = \log_2 M)$ by paying a suitable price in terms of the transmission bandwidth or the transmitted power. Figures 13.25 and 13.26 show that in multitone signaling the transmitted power decreases with M. However, the transmission bandwidth increases linearly with M, or exponentially with the rate increase factor k $(M = 2^k)$. Thus, multitone signaling is radically different from multiamplitude or multiphase signaling. In the latter, the bandwidth is independent of M, but the transmitted power increases as $M^2/\log_2 M = 2^{2k}/k$; that is, the power increases exponentially with the information-rate increase factor k. Thus, in the multitone case, the bandwidth increases exponentially with k, and in the multiamplitude or multiphase case, the power increases exponentially with k. Hence, we should use multiamplitude or multiphase signaling if the bandwidth is at a premium (as in telephone lines) and use multitone signaling when power is at a premium (as in space communication). A compromise exists between these two extremes. Let us investigate the possibility of increasing

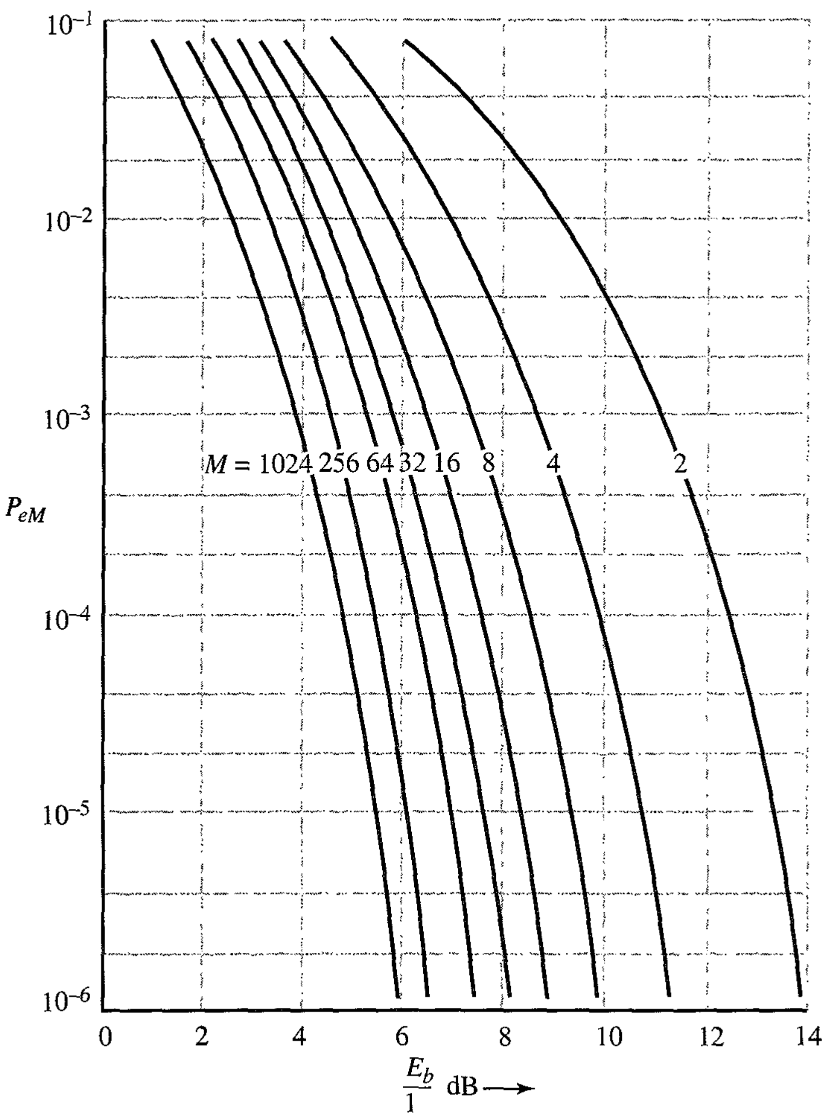


Figure 13.26 Error probability of noncoherent MFSK.

the information rate by a factor k simply by increasing the number of binary pulses transmitted by a factor k. In this case, the transmitted power increases linearly with k. Also because the bandwidth is proportional to the pulse rate, the transmission bandwidth increases linearly with k. Thus, in this case, we can increase the information rate by a factor of k by increasing both the transmission bandwidth and the transmitted power linearly with k, thus avoiding the phantom of the exponential increase that was required in the M-ary system. But here we must increase both the bandwidth and the power, whereas in the M-ary case the increase in information rate can be achieved by increasing either the bandwidth or the power. We have thus a great flexibility in trading various parameters and thus in our ability to match our resources to our requirements.

EXAMPLE 13.1 It is required to transmit 2.08×10^6 binary digits per second with $P_b \leq 10^{-6}$. Three possible schemes are considered:

(a) Binary

(b) 16-ary ASK
(c) 16-ary PSK

The channel noise PSD is $S_n(\omega) = 10^{-8}$. Determine the transmission bandwidth and the signal power required at the receiver input in each case.

(a) *Binary*: We shall consider polar signaling (the most efficient scheme),

$$P_b = P_e = 10^{-6} = Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right)$$

This yields $E_b/\mathcal{N} = 11.35$. The signal power $S_i = E_b R_b$. Hence,

$$S_i = 11.35\,\mathcal{N}R_b = 11.35(2 \times 10^{-8})(2.08 \times 10^6) = 0.47\,\text{W}$$

Assuming raised-cosine baseband pulses, the bandwidth B_T is

$$B_T = R_b = 2.08\,\text{MHz}$$

(b) *16-ary ASK*: Because each 16-ary symbol carries the information equivalent of $\log_2 16 = 4$ binary digits, we need to transmit only $R_M = (2.08 \times 10^6)/4 = 0.52 \times 10^6$ 16-ary pulses per second. This requires a bandwidth B_T of 520 kHz for baseband pulses and 1.04 MHz for modulated pulses (assuming raised-cosine pulses). Also,

$$P_b = 10^{-6} = \frac{P_{eM}}{\log_2 16}$$

Therefore,

$$P_{eM} = 4 \times 10^{-6} = 2\left(\frac{M-1}{M}\right) Q\left[\sqrt{\frac{6E_b \log_2 16}{\mathcal{N}(M^2-1)}}\right]$$

For $M = 16$, this yields $E_b = 0.499 \times 10^{-5}$. If the M-ary pulse rate is R_M, then

$$\begin{aligned} S_i &= E_{pM} R_M = E_b \log_2 M R_M \\ &= 0.499 \times 10^{-5} \times 4 \times (0.52 \times 10^6) = 9.34\,\text{W} \end{aligned}$$

(c) *16-ary PSK*: We need to transmit only $R_M = 0.52 \times 10^6$ pulses per second. For baseband pulses, this will require a bandwidth of 520 kHz. But PSK is a modulated signal, and the required bandwidth is $2(0.52 \times 10^6) = 1.04$ MHz. Also,

$$P_{eM} = 4P_b = 4 \times 10^{-6} \simeq 2Q\left[\sqrt{\frac{2\pi^2 E_b \log_2 16}{256\mathcal{N}}}\right]$$

This yields $E_b = 137.8 \times 10^{-8}$ and

$$\begin{aligned} S_i &= E_b \log_2 16 R_M \\ &= (137.8 \times 10^{-8}) \times 4 \times (0.52 \times 10^6) = 2.86\,\text{W} \end{aligned}$$

13.6 SYNCHRONIZATION

In synchronous, or coherent, detection, we need to achieve synchronization at three different levels: (1) carrier synchronization, (2) bit synchronization, and (3) word synchronization. For noncoherent detection, we need only the second and third levels of synchronization—which were discussed in Chapter 7. Here we shall consider only carrier synchronization.

Carrier synchronization is similar to bit synchronization, only more difficult. In bit synchronization, the problem is to achieve synchronism from bit interval to bit interval—which is of the order T_b. In carrier synchronization, we must achieve synchronism within a fraction of a cycle, and because the duration of one carrier cycle is $1/f_c \ll T_b$, the problem is severe. It should be remembered that the phase error θ that can be tolerated is much less than $\pi/2$. For example, if we are transmitting data at a rate of 2 Mbit/s, the bit interval is 0.5 μs. If this data is transmitted by PSK with a carrier frequency of 100 MHz, a phase of $\pi/2$ corresponds to 2.5 ns; that is, the synchronization must be achieved within an interval of much less than 2.5 ns!

Carrier synchronization is achieved by three general methods that are similar to those used for bit synchronization (see timing extraction in Sec. 7.5):

1. Using a primary or a secondary standard (i.e., transmitter and receiver slaved to a master timing source)
2. Transmitting a separate synchronizing signal (a pilot)
3. Self-synchronization, where the timing information is extracted from the received signal itself

The first method is expensive and is suitable only for large data systems, not for point-to-point systems.

The second method uses part of the channel capacity to transmit timing information and causes some degradation in performance (see Prob. 13.2-1). But this is a widely used method for point-to-point communication systems. A pilot may be transmitted by frequency-division multiplexing (by choosing a pilot of frequency at which the signal PSD has a null) or by time-division multiplexing (in which the modulated signal is interrupted for a short period of time, during which the synchronizing signal is transmitted).

A baseband signaling scheme with a dc null—such as bipolar, duobinary, or split-phase—is preferred, because such signals after modulation have a spectral null at the carrier frequency. This facilitates the separation of the pilot at the receiver.

The self-synchronization method extracts the carrier by squaring the incoming signal or by using a Costas loop, as discussed in Sec. 4.7. But because these methods yield sign ambiguities, they cannot be used for PSK unless differential coding is used.

REFERENCES

1. S. Pasupathy, "Minimum Shift Keying: A Spectrally Efficient Modulation," *IEEE Commun. Soc. Mag.*, vol. 17, pp. 14–22, July 1979.
2. J. J. Spilker, *Digital Communications by Satellite*, Prentice-Hall, Englewood Cliffs, NJ, 1977.
3. A. J. Viterbi, *Principles of Coherent Communication*, McGraw-Hill, New York, 1966.

4. J. M. Wozenraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965, chap. 7.
5. J. I. Marcum, "Statistical Theory of Target Detection by Pulsed Radar," *IRE Trans. Inf. Theory*, vol. IT-6, pp. 259–267, April 1960.
6. M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, McGraw-Hill, New York, 1966.
7. D. V. Sarwate and M. B. Pursley, "Crosscorrelation Properties of Pseudorandom and Related Sequences," *Proc. IEEE*, vol. 68, pp. 593–619, May 1980.
8. M. B. Pursley, "Performance Analysis for Phase-Coded Spread-Spectrum Multiple-Access Communication—Part I: System Analysis," *IEEE Trans. Commun.*, vol. COM-25, pp. 795–799, Aug. 1977.
9. E. Geraniotis and M. B. Pursley, "Error Probabilities for Direct Sequence Spread Spectrum Multiple Access Communications—Part II: Approximations," *IEEE Trans. Commun.*, vol. COM-30, pp. 985–995, May 1982.
10. J. S. Lehnert, "An Efficient Technique for Evaluating Direct-Sequence Spread-Spectrum Communications," *IEEE Trans. Commun.*, vol. 37, pp. 851–858, Aug. 1989.
11. R. K. Morrow and J. S. Lehnert, "Bit-to-Bit Error Dependence in Slotted DS/SSMA Packet Systems with Random Signal Sequences," *IEEE Trans. Commun.*, vol. COM-37, pp. 1052–1061, Oct. 1989.
12. J. M. Holtzman, "A Simple Accurate Method to Calculate Spread-Spectrum Multiple Access Error Probabilities," *IEEE Trans. Commun.*, vol. COM-40, pp. 461–464, March 1992.
13. B. D. Woerner and R. Cameron, "An Analysis of CDMA with Imperfect Power Control," *Proc. 42nd IEEE Vehicular Technology Conf.*, pp. 977–980, May 1992.
14. D. L. Schilling et al., "Broadband CDMA for Personal Communications Systems," *IEEE Commun. Mag.*, vol. 29, pp. 86–93, Nov. 1991.
15. G. L. Turin, "Introduction to Spread-Spectrum Antimultipath Techniques and Their Application to Urban Digital Radio," *Proc. IEEE*, vol. 68, pp. 328–353, March 1980.
16. R. Price and P. E. Green, Jr., "A Communication Technique for Multipath Channels," *Proc. IRE*, vol. 46, pp. 555–570, March 1958.
17. E. O. Geraniotis and M. B. Pursley, "Error Probabilities for Slow-Frequency-Hopped Spread-Spectrum Multiple-Access Communications Over Fading Channels," *IEEE Transactions on Communications*, vol. Com-30, no. 5, pp 996–1009, May 1982.
18. S. G. Wilson, *Digital Modulation and Coding*, Prentice Hall, Upper Saddle River, NJ, 1996.

PROBLEMS

13.1-1 The so-called integrate-and-dump filter is shown in Fig. P13.1-1. The feedback amplifier is an ideal integrator. The switch s_1 closes momentarily and then opens at the instant $t = T_b$, thus dumping all the charge on C and causing the output to go to zero. The switch s_2 samples the output immediately before the dumping action.

(**a**) Sketch the output $p_o(t)$ when a square pulse $p(t)$ is applied to the input of this filter.

(**b**) Sketch the output $p_o(t)$ of the filter matched to the square pulse $p(t)$.

(**c**) Show that the performance of the integrate-and-dump filter is identical to that of the matched filter; that is, show that ρ in both cases is identical.

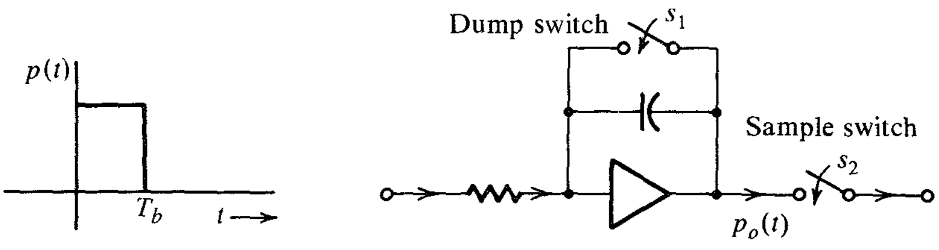
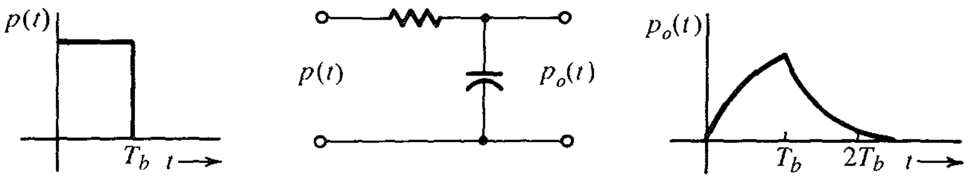


Figure P13.1-1

13.1-2 An alternative to the optimum filter is a suboptimum filter, where we assume a particular filter form and adjust its parameters to maximize ρ. Such filters are inferior to the optimum filter but may be simpler to design.

For a rectangular pulse $p(t)$ of height A and width T_b at the input (Fig. P13.1-2), determine $\rho_{\max}$ if, instead of the matched filter, a one-stage RC filter with $H(\omega) = 1/(1 + j\omega RC)$ is used. Assume a white gaussian noise of PSD $\mathcal{N}/2$. Show that the optimum performance is achieved when $1/RC = 1.26/T_b$. *Hint*: Set $d\rho^2/dx = 0\ (x = T_b/RC)$.



(a) (b) (c)

Figure P13.1-2 Suboptimum filter.

13.2-1 As discussed in Sec. 13.3, in coherent schemes, a small pilot is added for synchronization. Because the pilot does not carry information, it causes degradation in P_b. Consider coherent PSK using the following two pulses of duration T_b each:

$$p(t) = A\sqrt{1 - m^2}\ \cos \omega_c t + Am\ \sin \omega_c t$$

$$q(t) = -A\sqrt{1 - m^2}\cos \omega_c t + Am\ \sin \omega_c t$$

where $Am\ \sin \omega_c t$ is the pilot. Show that when the channel noise is white gaussian,

$$P_b = Q\left[\sqrt{\frac{2E_b(1 - m^2)}{\mathcal{N}}}\right]$$

Hint: Use Eq. (13.18b).

13.2-2 For polar binary communication systems, each error in the decision has some cost. Suppose that when m = **1** is transmitted and we read it as m = **0** at the receiver, a quantitative penalty, or cost, C_{10} is assigned to such an error, and, similarly, a cost C_{01} is assigned when m = **0** is transmitted and we read it as m = **1**. For the polar case where $P_m(\mathbf{0}) = P_m(\mathbf{1}) = 0.5$, show that for white gaussian channel noise the optimum threshold that minimizes the overall cost is not 0 but is a_o, given by

$$a_o = \frac{\mathcal{N}}{4} \ln \frac{C_{01}}{C_{10}}$$

Hint: See Hint for Prob. 10.2-11.

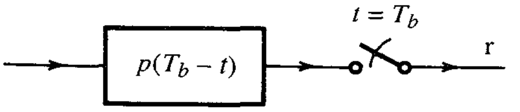
13.2-3 For a polar binary system with unequal message probabilities, show that the optimum decision threshold a_o is given by

$$a_o = \frac{\mathcal{N}}{4} \ln \frac{P_m(\mathbf{0})C_{01}}{P_m(\mathbf{1})C_{10}}$$

where C_{01} and C_{10} are the cost of the errors as explained in Prob. 13.2-2, and $P_m(\mathbf{0})$ and $P_m(\mathbf{1})$ are the probabilities of transmitting **0** and **1**, respectively. *Hint*: See Hint for Prob. 10.2-11.

13.5-1 For 3-ary communication, messages are chosen from any one of three symbols, m_{-1}, m_0, and m_1, which are transmitted by pulses $-p(t)$, 0, and $p(t)$, respectively. A filter matched to $p(t)$ is used at the receiver. If r is the matched filter output at T_M, plot $p_r(r|m_i)$ ($i = -1$, 0, and 1), and if $P(m_{-1}) = P(m_0) = P(m_1)$, determine the optimum decision thresholds and the error probability P_e. The energy of the pulse $p(t)$ is E_p and the channel noise PSD is $S_n(\omega) = \mathcal{N}/2$.

13.5-2 Binary data is transmitted by using a pulse $p(t)$ for **0** and a pulse $3p(t)$ for **1**. Show that the optimum receiver for this case is a filter matched to $p(t)$ with a detection threshold $2E_p$, as shown in Fig. P13.5-2. Determine the error probability P_b of this receiver as a function of $E_b/\mathcal{N}$ if **0** and **1** are equiprobable.



Decision:
0 if $r < 2E_p$
1 if $r > 2E_p$

Figure P13.5-2

13.5-3 A binary source emits data at a rate of 256,000 bit/s. Multiamplitude-shift keying (MASK) with $M = 2$, 16, and 32 is considered. In each case, determine the signal power required at the receiver input and the minimum transmission bandwidth required if $S_n(\omega) = 10^{-8}$ and the bit error rate P_b, is required to be less than 10^{-7}.

13.5-4 Repeat Prob. 13.5-3 for M-ary PSK.

14 OPTIMUM SIGNAL DETECTION

Our scope of the discussion of signal detection in Chapter 13 was rather limited. In the first place, we restricted ourselves primarily to binary systems with a receiver of assumed (linear) form. Although we optimized the linear type of receiver, we have no assurance that there does not exist another type of receiver that might be superior to the optimum linear receiver derived in Chapter 13. Second, we did not derive the optimum linear receiver for the M-ary case in general. In Chapter 13 we only analyzed a few M-ary schemes with an assumed receiver structure.

In this chapter, we shall analyze the problem of digital signal detection from a more fundamental point of view. We shall determine the optimum receiver (in the sense of minimizing the error probability) for general M-ary signaling in the presence of additive white gaussian noise (AWGN). We shall place no constraints on the receiver. Rather, we shall try to answer the question: What receiver will yield the minimum error probability?

Such an analysis is greatly facilitated by a geometrical representation of signals. We shall now show that a signal is in reality an n-dimensional vector and can be represented by a point in an n-dimensional hyperspace. The foundations for such a viewpoint were laid in Chapter 2. Here we shall refine and concretize our argument. The reader may want to review the concepts of vector representation of a signal in Chapter 2.

14.1 GEOMETRICAL REPRESENTATION OF SIGNALS: SIGNAL SPACE

We are used to 3-dimensional physical space and 3-dimensional vectors in this space. There is no reason, however, to restrict ourselves to three dimensions only. We can extend the concept to an n-dimensional space, although it may be hard to visualize the space for $n > 3$.

To begin with, we note that the familiar 3-dimensional vectors are nothing but entities specified by three numbers $(x_1,\ x_2,\ x_3)$ in a certain order (ordered 3-tuple)—nothing more, nothing less. Extending this concept, we say that any entity specified by n numbers in a certain order (ordered n-tuple) is an n-dimensional vector. Thus, if an entity is specified by an ordered n-tuple $(x_1, x_2, \ldots, x_n)$, it is an n-dimensional vector $\mathbf{x}$. We define n unit vectors $\varphi_1, \varphi_2, \ldots, \varphi_n$ as

$$\begin{aligned}\varphi_1 &= (1, 0, 0, \ldots, 0)\\ \varphi_2 &= (0, 1, 0, \ldots, 0)\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ \varphi_n &= (0, 0, 0, \ldots, 1)\end{aligned} \tag{14.1}$$

Any vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ can be expressed as a linear combination of n unit vectors,

$$\mathbf{x} = x_1\varphi_1 + x_2\varphi_2 + \cdots + x_n\varphi_n \tag{14.2a}$$

$$= \sum_{k=1}^{n} x_k\varphi_k \tag{14.2b}$$

We define the scalar product $\mathbf{x} \cdot \mathbf{y}$ as

$$\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k \tag{14.3}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is another n-dimensional vector in the same space. Vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be **orthogonal** if

$$\mathbf{x} \cdot \mathbf{y} = 0 \tag{14.4}$$

The **length** of a vector $\mathbf{x}$ is $|\mathbf{x}|$, defined by

$$|\mathbf{x}|^2 = \mathbf{x} \cdot \mathbf{x} = \sum_{k=1}^{n} x_k^2 \tag{14.5}$$

A set of n-dimensional vectors is said to be independent if none of the vectors in this set can be represented as a linear combination of the remaining vectors in the set. Thus, if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ is an independent set, then it is impossible to find constants $a_1, a_2, \ldots, a_m$ (not all zero) such that

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_m\mathbf{x}_m = 0 \tag{14.6}$$

An n-dimensional space can have at most n independent vectors. If a space has a maximum of n independent vectors, then every vector $\mathbf{x}$ in this space can be expressed as a linear combination of these n independent vectors. If not, then $\mathbf{x}$ becomes a member of the independent vector space. But this is not possible with the assumption that there exists a maximum of n independent vectors. Thus, any vector in this space can be specified by n numbers and, hence, is an n-dimensional space. It also follows from this discussion that the dimensionality of a space can be, at most, as large as the total number of vectors in the space.

A subset of vectors in a given n-dimensional space can have dimensionality less than n. Thus, in a 3-dimensional space, all vectors lying in one plane can be specified by two dimensions, and all vectors lying along a line can be specified by one dimension.

A set of n independent vectors in an n-dimensional space are called **basis vectors** because every vector in this space can be expressed as a linear combination of these n independent vectors. A set of basis vectors form coordinate axes, and they are not unique. The n unit vectors in Eq. (14.1) are independent and can serve as basis vectors. These vectors have an additional property in that they are all mutually **orthogonal**, that is,

$$\varphi_j \cdot \varphi_k = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases} \tag{14.7}$$

Such a set is an **orthonormal** set of vectors because in addition to being orthogonal, their lengths are unity.

A vector $\mathbf{x}$ $(x_1,\ x_2, \ldots,\ x_n)$ can be represented as

$$\mathbf{x} = x_1\varphi_1 + x_2\varphi_2 + \cdots x_n\varphi_n$$

If we know $\mathbf{x}$, its components x_k can be found by taking the scalar product of both sides with φ_k. Using Eq. (14.7) we get

$$\mathbf{x} \cdot \varphi_k = x_k \qquad k = 1,\ 2, \ldots,\ n \tag{14.8}$$

One final observation: in order to represent an n-dimensional vector completely, we need n basis vectors (independent vectors). In general, an n-dimensional vector cannot be completely represented by less than n basis vectors of the space. A set of n basis vectors forms a **complete set**.

Signal as a Vector

Any entity that can be represented by an n-tuple is an n-dimensional vector. If a signal $x(t)$ can be specified by an n-tuple, it, too, is a vector. As we know, a signal $x(t)$ band-limited to B Hz can be specified by the values of its Nyquist samples. Consequently, such a signal, too, is a vector. We can express $x(t)$ as [see Eq. (6.10)]

$$x(t) = \sum_k x\left(\frac{k}{2B}\right) \text{ sinc } (2\pi Bt - k\pi) \tag{14.9}$$

Thus, $x(t)$ can be specified by an n-tuple (its Nyquist sample values). With this introduction, let us now turn to a systematic development of the signal-space concept.

We define n signals $\varphi_1(t),\ \varphi_2(t), \ldots,\ \varphi_n(t)$ as independent if none of these n signals can be represented by a linear combination of the remaining $n-1$ signals. This means it is impossible to find constants $a_1,\ a_2,\ \ldots,\ a_n$ (not all zero) such that

$$a_1\varphi_1(t) + a_2\varphi_2(t) + \cdots + a_n\varphi_n(t) = 0 \tag{14.10}$$

Suppose that a signal $x(t)$ can be represented by a linear combination of n independent signals $\{\varphi_k(t)\}$ as

$$x(t) = x_1\varphi_1(t) + x_2\varphi_2(t) + \cdots + x_n\varphi_n(t) \tag{14.11a}$$

$$= \sum_{k=1}^{n} x_k\varphi_k(t) \tag{14.11b}$$

If every signal in a certain signal space can be represented by a linear combination of n independent signals $\{\varphi_k(t)\}$, then we have an n-dimensional signal space.

The signal set $\{\varphi_k(t)\}$ will be assumed to be *orthogonal*, that is,*

$$\int_{-\infty}^{\infty} \varphi_j(t)\varphi_k(t)\,dt = \begin{cases} 0 & j \neq k \\ k_j & j = k \end{cases} \tag{14.12}$$

If $k_j = 1$ for all j, then the set is *orthonormal*. For an orthonormal set, the coefficients x_k in Eq. (14.11) can be obtained by multiplying both sides of Eq. (14.11) by $\varphi_k(t)$, then integrating and using Eq. (14.12) (with $k_j = 1$):

$$x_k = \int_{-\infty}^{\infty} x(t)\varphi_k(t)\,dt \tag{14.13}$$

Once the basis signals $\{\varphi_k(t)\}$ are specified, we can represent a signal $x(t)$ by an n-tuple $(x_1, x_2, \ldots, x_n)$. Alternately we may represent this signal geometrically by a point $(x_1, x_2, \ldots, x_n)$ in an n-dimensional space. We can now associate a vector $\mathbf{x}$ $(x_1, x_2, \ldots, x_n)$ with the signal $x(t)$. Note that the basis signal $\varphi_1(t)$ is represented by the corresponding basis vector $\boldsymbol{\varphi}_1(1, 0, 0, \ldots, 0)$, and $\varphi_2(t)$ is represented by $\boldsymbol{\varphi}_2(0, 1, 0, \ldots, 0)$, and so on.

Turning our attention to Eq. (14.9), we see that in this case the basis signals are $\{\text{sinc}\,(2\pi Bt - k\pi)\}$. This set is orthogonal because (see Prob. 3.7-4)

$$\int_{-\infty}^{\infty} \text{sinc}\,(2\pi Bt - j\pi)\,\text{sinc}\,(2\pi Bt - k\pi)\,dt = \begin{cases} 0 & j \neq k \\ 1/2B & j = k \end{cases} \tag{14.14}$$

We can normalize this set by multiplying each member by $\sqrt{2B}$. Thus, if we define

$$\varphi_k(t) = \sqrt{2B}\,\text{sinc}\,(2\pi Bt - k\pi) \tag{14.15}$$

then

$$x(t) = \sum_k x_k \varphi_k(t) \tag{14.16a}$$

where

$$x_k = \frac{1}{\sqrt{2B}}\,x\left(\frac{k}{2B}\right) \tag{14.16b}$$

and $\{\varphi_k(t)\}$ is an orthonormal set of signals. Thus, any band-limited signal $x(t)$ can be represented by a point

$$(\ldots, x_{-k}, \ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots, x_k, \ldots)$$

where x_k is the kth Nyquist sample of $x(t)$ divided by $1/\sqrt{2B}$. To determine the dimensionality of this signal space, recall that a band-limited signal cannot be time-limited (i.e., it exists over an infinite time interval). Hence, the total number of samples at a Nyquist rate of $2B$ samples

---

* If $\{\varphi_k(t)\}$ is complex, orthogonality implies

$$\int_{-\infty}^{\infty} \varphi_j(t)\varphi_k^*(t)\,dt = 0$$

and Eq. (14.13) becomes

$$x_k = \int_{-\infty}^{\infty} x(t)\varphi_k^*(t)\,dt$$

per second will be infinite, and the dimensionality is infinite. Higher dimensions, however, can be ignored, because their contribution is negligible. For example, a band-limited signal with a finite energy may exist over an infinite time interval. But its amplitude for large values of t must approach 0, or the energy

$$E_x = \int_{-\infty}^{\infty} x^2(t)\, dt$$

will not be finite. Hence, the amplitude of every band-limited signal is negligible beyond some $|t| = T/2$. Such a signal is essentially time-limited to T seconds. We can argue in the same way that a signal time-limited to T seconds is essentially band-limited to B Hz. Because the Nyquist rate is $2B$ samples per second, the total number of samples over T seconds is* $2BT+1$, where T and B are interpreted as before. If $x(t)$ is band-limited to B, then T is its essential time duration, and if $x(t)$ is time-limited to T, then it is essentially band-limited to B Hz. A rigorous development of this result, as well as an estimation of the error in ignoring higher dimensions (those beyond $2BT + 1$), can be found in Landau and Pollak.[1]

Just as there are an infinite number of possible sets of basis vectors for a vector space, there are an infinite number of possible sets of basis signals for a given signal space. For a band-limited signal space, $\{\text{sinc}\,(2\pi Bt - k\pi)\}$ is one possible set of basis signals.

Scalar Product

In a certain signal space, let $x(t)$ and $y(t)$ be two signals represented by vectors $\mathbf{x}\,(x_1, x_2, \ldots, x_n)$ and $\mathbf{y}\,(y_1, y_2, \ldots, y_n)$. If $\{\varphi_k(t)\}$ are the orthonormal basis signals, then

$$x(t) = \sum_i x_i \varphi_i(t)$$

$$y(t) = \sum_j y_j \varphi_j(t)$$

Hence,

$$\int_{-\infty}^{\infty} x(t)y(t)\, dt = \int_{-\infty}^{\infty} \left[\sum_i x_i \varphi_i(t)\right] \left[\sum_j y_j \varphi_j(t)\right] dt$$

Because the basis signals are orthonormal, from Eq. (14.12) with $k_j = 1$, we obtain

$$\int_{-\infty}^{\infty} x(t)y(t)\, dt = \sum_k x_k y_k \tag{14.17a}$$

The right-hand side of Eq. (14.17a), however, is by definition the scalar product of vectors $\mathbf{x}$ and $\mathbf{y}$;

$$\mathbf{x} \cdot \mathbf{y} = \sum_k x_k y_k$$

Hence, we have

$$\mathbf{x} \cdot \mathbf{y} = \int_{-\infty}^{\infty} x(t)y(t)\, dt \tag{14.17b}$$

---

* Including end samples.

If $x(t)$ and $y(t)$ are mutually orthogonal, then it follows from Eq. (14.17b) that the corresponding vectors $\mathbf{x}$ and $\mathbf{y}$ are also orthogonal. We conclude that the integral of the product of two signals is equal to the scalar product of the corresponding vectors.

Energy of a Signal

For a signal $x(t)$, the energy E_x is given by

$$E_x = \int_{-\infty}^{\infty} x^2(t)\, dt$$

It follows from Eq. (14.17b) that

$$E_x = \mathbf{x} \cdot \mathbf{x} = |\mathbf{x}|^2 \tag{14.18}$$

where $|\mathbf{x}|^2$ is the square of the length of the vector $\mathbf{x}$. Hence, the signal energy is given by the square of the length of the corresponding vector.

EXAMPLE 14.1 A signal space consists of four signals $s_1(t)$, $s_2(t)$, $s_3(t)$, and $s_4(t)$, as shown in Fig. 14.1. Determine a suitable set of basis vectors and the dimensionality of the signals. Represent these signals geometrically in the vector space.

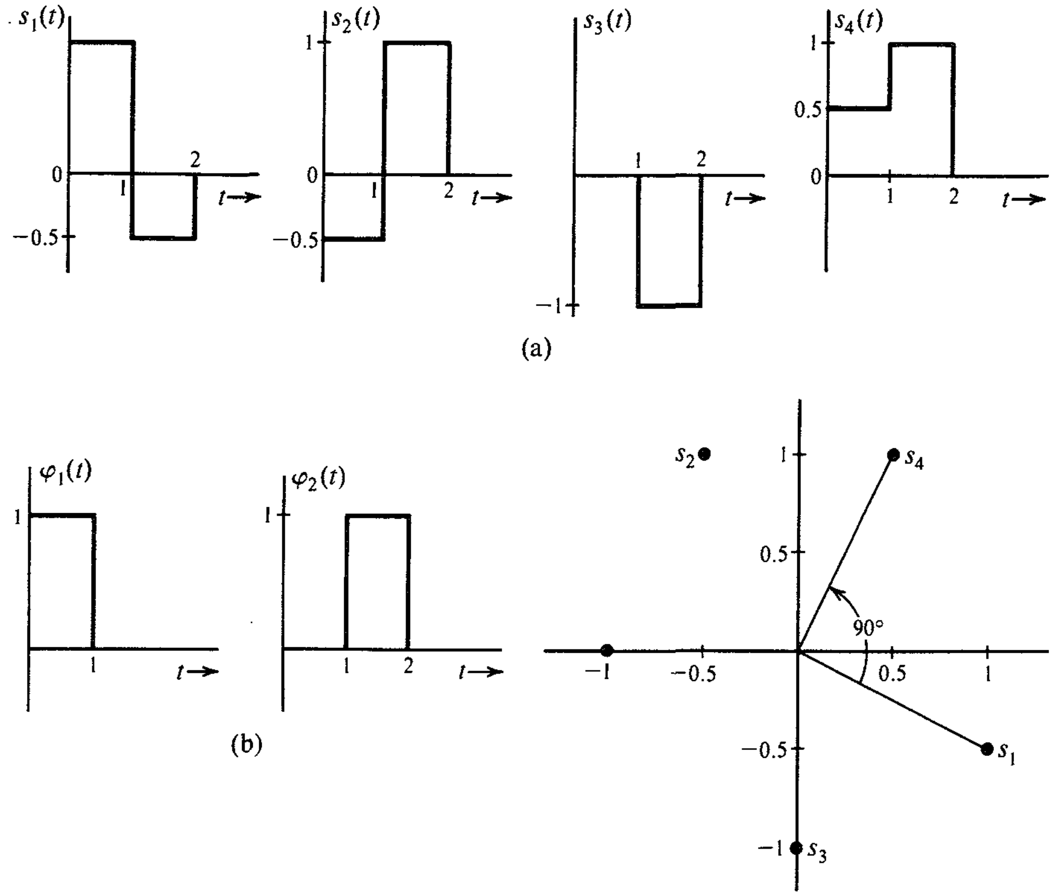

$s_1(t)$
0
-0.5
1
2
$t \rightarrow$
$s_2(t)$
1
0
-0.5
1
2
$t \rightarrow$
$s_3(t)$
1
2
-1
$t \rightarrow$
$s_4(t)$
1
0.5
0
1
2
$t \rightarrow$
(a)
$\varphi_1(t)$
1
1
$t \rightarrow$
$\varphi_2(t)$
1
1
2
$t \rightarrow$
(b)
s_2
1
s_4
0.5
90°
-1
-0.5
0.5
1
-0.5
s_1
-1
s_3

Figure 14.1 Signals and their representation in signal space

The two rectangular pulses $\varphi_1(t)$ and $\varphi_2(t)$ in Fig. 14.1b are suitable as a basis signal set. In terms of this set, the vectors s_1, s_2, s_3, and s_4 corresponding to signals $s_1(t)$, $s_2(t)$, $s_3(t)$, and $s_4(t)$ are $s_1 = (1, \ -0.5)$, $s_2 = (-0.5, \ 1)$, $s_3 = (0, \ -1)$, and $s_4 = (0.5, \ 1)$. These points are plotted in Fig. 14.1c. Observe that

$$s_1 \cdot s_4 = 0.5 - 0.5 = 0$$

Hence, s_1 and s_4 are orthogonal. This result may be verified from the fact that

$$\int_{-\infty}^{\infty} s_1(t)s_4(t)\, dt = 0$$

Note that each point in the signal space in Fig. 14.1c corresponds to some waveform.

Systematic Determination of an Orthogonal Basis Set

We have shown that the dimensionality of a vector space is equal to the maximum number of independent vectors in the space. Thus, in an n-dimensional space, there can be no more than n vectors that are independent. Alternatively, it is always possible to find a set of n vectors that are independent. Once such a set (basis set) is chosen, any vector in this space can be expressed in terms of (as a linear combination of) the vectors in this set. This set of n independent vectors is by no means unique. The reader is familiar with this fact in the physical space of three dimensions, where one can find an infinite number of independent sets of three vectors, each forming a valid coordinate system. If we are given a set of n independent vectors, it is possible to obtain from this set another set of n independent vectors that is orthogonal. This is done by the **Gram-Schmidt orthogonalization process** discussed in Appendix C.

A deterministic signal can be represented by one point in a signal space. A random process, on the other hand, consists of an ensemble of waveforms, each of which maps into a point in a signal space. Hence, a random process appears as an ensemble of points in a signal space. Let us consider in detail the gaussian random process and its representation in a signal space.

14.2 GAUSSIAN RANDOM PROCESS

In order to specify a gaussian random process, we need to familiarize ourselves first with jointly gaussian random variables.

Jointly Gaussian Random Variables

Random variables (RVs) $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$ are said to be jointly gaussian if their joint PDF is given by

$$p_{\mathrm{x}_1\mathrm{x}_2\ldots\mathrm{x}_n}(x_1, x_2, \ldots, x_n)$$
$$= \frac{1}{(2\pi)^{n/2}\sqrt{|K|}} \exp\left[\frac{-1}{2|K|}\sum_i\sum_j \Delta_{ij}(x_i - \overline{\mathrm{x}_i})(x_j - \overline{\mathrm{x}_j})\right] \tag{14.19}$$

where K is the covariance matrix

$$K = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} \tag{14.20a}$$

and the covariance of x_i and x_j is

$$\sigma_{ij} = \overline{(x_i - \bar{x}_i)(x_j - \bar{x}_j)} \tag{14.20b}$$

$|K|$ is the determinant of matrix K, and Δ_{ij} is the cofactor for the element σ_{ij}. Note that $\sigma_{ii} = \sigma_i^2$.

Gaussian variables are important not only because they are frequently observed, but also because they have certain properties that simplify many mathematical operations that are practically impossible or very difficult for other types of RVs. These properties are as follows:

Property 1: The gaussian density is completely specified by only the first and second moments (means and covariances). This follows from Eq. (14.19). It can be shown that, in general, the probability density of an RV depends on all the moments $\overline{x^n}$ ($n = 1, 2, \ldots$) of the variable. A gaussian variable is a special case where only the first two moments (means) and the second moments (covariances) are necessary to specify the distribution.

Property 2: If n jointly gaussian variables $x_1, x_2, \ldots, x_n$ are uncorrelated, they are independent.

If the n variables are uncorrelated, $\sigma_{ij} = 0$ ($i \neq j$), and Eq. (14.19) reduces to

$$p_{x_1x_2\cdots x_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{-(x_i - \overline{x_i})^2}{2\sigma_i^2}\right] \tag{14.21a}$$

$$= p_{x_1}(x_1)p_{x_2}(x_2)\ldots p_{x_n}(x_n) \tag{14.21b}$$

As we observed earlier, independent variables are always uncorrelated, but uncorrelated variables are not necessarily independent. For the case of jointly gaussian RVs, however, uncorrelatedness implies independence.

Property 3: When $x_1, x_2, \ldots, x_n$ are jointly gaussian, all the marginal densities, such as $p_{x_i}(x_i)$, and all the conditional densities, such as $p_{x_ix_j|x_kx_l\cdots x_p}(x_i, x_j|x_k, x_l, \ldots, x_p)$, are gaussian. This property can be readily verified (see Prob. 10.2-9).

Property 4: Linear combinations of jointly gaussian variables are also jointly gaussian. Thus, if we form m variables $y_1, y_2, \ldots, y_m$ ($m \leq n$) such that

$$y_i = \sum_{k=1}^{n} a_{ik}x_k$$

then $y_1, y_2, \ldots, y_m$ are also jointly gaussian variables.

Definition of a Gaussian Random Process

A random process $x(t)$ is said to be Gaussian if the RVs $x(t_1), x(t_2), \ldots, x(t_n)$ are jointly gaussian [Eq. (14.19)] for every n and for every set $(t_1, t_2, \ldots, t_n)$.

For convenience, the RV $x(t_i)$ will be denoted by x_i. Hence, the joint PDF of RVs $x_1, x_2, \ldots, x_n$ of a gaussian random process is given by Eq. (14.19). Observe that

$$\begin{aligned}\sigma_{ij} &= \overline{(x_i - \bar{x}_i)(x_j - \bar{x}_j)} \\ &= \overline{x_i x_j} - \bar{x}_i \bar{x}_j - \bar{x}_i \bar{x}_j + \bar{x}_i \bar{x}_j \\ &= \overline{x_i x_j} - \bar{x}_i \bar{x}_j \\ &= \overline{x(t_i) x(t_j)} - \bar{x}_i \bar{x}_j \\ &= R_x(t_i, t_j) - \overline{x(t_i)}\,\overline{x(t_j)} \qquad (14.22)\end{aligned}$$

This shows that a gaussian random process is completely specified by its autocorrelation function $R_x(t_i, t_j)$ and its mean value $\overline{x(t)}$.

Stationary Gaussian Random Process

Our discussion of the gaussian process thus far applies to stationary and nonstationary processes. We have shown that the gaussian process is completely specified by its autocorrelation function $R_x(t_i, t_j)$ and its mean value function.* An important corollary of this statement is that if the autocorrelation function $R_x(t_i, t_j)$ and the means are unaffected by a shift of the time origin, then all of the statistics of the process are unaffected by a shift of the time origin. In other words, the process is stationary. Thus, if

$$R_x(t_i, t_j) = R_x(t_i - t_j) \qquad (14.23)$$

and

$$\overline{x(t)} = \text{constant for all } t$$

the process is stationary if it is gaussian. But the condition in Eq. (14.23) defines wide-sense stationarity. Hence, *for a gaussian process, wide-sense stationarity implies stationarity in the strict sense.*

For a stationary gaussian process, the covariance σ_{ij} becomes

$$\sigma_{ij} = R_x(t_j - t_i) - \bar{x}^2 \qquad (14.24)$$

We shall once again stress the point that distinguishes the gaussian from the nongaussian process. The complete statistics of a gaussian process are determined from its autocorrelation function (which is a second-order parameter) and its mean value. This is the property that simplifies the study of gaussian processes. In general, for nongaussian processes higher order statistics cannot be determined from lower order statistics. For the gaussian process, however, the mean and the second-order parameter σ_{ij} determine all the higher order statistics. This very property also enables us to state that a gaussian process is strictly stationary if it is wide-sense stationary.

Transmission of a Gaussian Process through a Linear System

Another significant property of the gaussian process is that the response of a linear system to a gaussian process is also a gaussian process. This can be shown as follows. Let $x(t)$ be a

---

[*] For nonstationary processes (or processes that are not wide-sense stationary), the mean value is a function of t.

gaussian process applied to the input of a linear system whose unit impulse response is $h(t)$. If $\mathrm{y}(t)$ is the output (response) process, then

$$\begin{aligned}\mathrm{y}(t) &= \int_{-\infty}^{\infty} \mathrm{x}(t-\tau)h(\tau)\,d\tau \\ &= \lim_{\Delta\tau \to 0} \sum_{k=-\infty}^{\infty} \mathrm{x}(t-\tau_k)h(\tau_k)\,\Delta\tau \qquad \tau_k = k\Delta\tau\end{aligned}$$

Because $\mathrm{x}(t)$ is a gaussian process, all the variables $\mathrm{x}(t-\tau_k)$ are jointly gaussian (by definition). Hence, the variables $\mathrm{y}(t_1),\ \mathrm{y}(t_2),\ \ldots,\ \mathrm{y}(t_n)$ for all n and every set $(t_1,\ t_2,\ \ldots,\ t_n)$ are linear combinations of variables that are jointly gaussian. Therefore, the variables $\mathrm{y}(t_1),\ \mathrm{y}(t_2),\ \ldots,\ \mathrm{y}(t_n)$ must be jointly gaussian, according to the earlier discussion. It follows that the process $\mathrm{y}(t)$ is a gaussian process.

To summarize, the gaussian random process has the following properties:

1. A gaussian random process is completely specified by its autocorrelation function and mean value.
2. If a gaussian random process is wide-sense stationary, then it is stationary in the strict sense.
3. The response of a linear system to a gaussian random process is also a gaussian random process.

Geometrical Representation of a Random Process

Let $\mathrm{x}(t)$ be a random process and $\varphi_1(t),\ \varphi_2(t),\ \ldots,\ \varphi_n(t)$ a complete set of orthonormal basis signals for this space. We can express $\mathrm{x}(t)$ as

$$\begin{aligned}\mathrm{x}(t) &= \mathrm{x}_1\varphi_1(t) + \mathrm{x}_2\varphi_2(t) + \cdots + \mathrm{x}_n\varphi_n(t) \\ &= \sum_{k=1}^{n} \mathrm{x}_k\varphi_k(t) \qquad (14.25)\end{aligned}$$

where

$$\mathrm{x}_k = \int_{-\infty}^{\infty} \mathrm{x}(t)\varphi_k(t)\,dt \qquad (14.26)$$

Note that $\mathrm{x}(t)$ is a random process, consisting of an ensemble of sample functions. The coefficients x_k in Eq. (14.25) will be different for each sample function. Consequently, $\mathrm{x}_1,\ \mathrm{x}_2, \ldots,\ \mathrm{x}_n$ appearing in Eq. (14.25) are RVs. Each sample function will have a specific set $(x_1,\ x_2,\ \ldots,\ x_n)$ and will map into one point in the signal space. Hence, the ensemble of sample functions will map into an ensemble of points in the signal space, as shown in Fig. 14.2. This figure shows only a 3-dimensional graph because it is not possible to show a higher dimensional one. But it is sufficient to indicate the idea.

Each time the random experiment is repeated, the outcome (the sample function) is a certain point $\boldsymbol{x}$. The ensemble of points in the signal space appears as a cloud, with the density of points directly proportional to the probability of observing $\mathbf{x}$ in that region. If we denote the joint PDF of $\mathrm{x}_1,\ \mathrm{x}_2,\ \ldots,\ \mathrm{x}_n$ by $p_{\mathbf{x}}(\boldsymbol{x})$, then

$$p_{\mathbf{x}}(\boldsymbol{x}) = p_{\mathrm{x}_1\mathrm{x}_2\cdots\mathrm{x}_n}(x_1,\ x_2,\ \ldots,\ x_n)$$

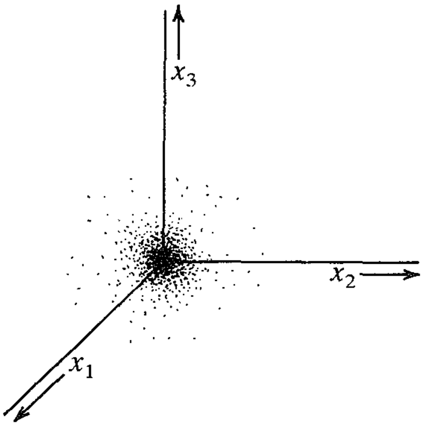


Figure 14.2 Geometrical representation of a gaussian random process.

Thus, $p_{\mathrm{x}}(\boldsymbol{x})$ has a certain value at each point in the signal space, and $p_{\mathrm{x}}(\boldsymbol{x})$ represents the relative probability (cloud density) of observing $\mathbf{x} = \boldsymbol{x}$.

A Word about Notation: Let us clarify the notation used here to avoid confusion later. As before, we use roman type to denote an RV or a random process. Thus, x or x(t) represents an RV or a random process. A particular value assumed by the RV in a certain trial is denoted by italic type. Thus, x represents the value assumed by x. Similarly, $x(t)$ represents a particular sample function of the random process x(t). In the case of random vectors, we follow the same convention; a random vector is denoted by roman boldface type, and a particular value assumed by the vector in a certain trial is represented by boldface italic type. Thus, **r** denotes a random vector representing a random process r(t), but $\boldsymbol{r}$ is a particular value of **r** and represents a particular received waveform (sample function) $r(t)$ in some trial. Note that roman type represents random entities and italic type represents particular values (which are, of course, nonrandom).

White Gaussian Noise

Consider a white noise process $\mathrm{n}_w(t)$ with PSD $\mathcal{N}/2$. Let $\varphi_1(t),\ \varphi_2(t),\ \ldots$ be a complete set of orthonormal basis signals for this space. We can express $\mathrm{n}_w(t)$ as

$$\begin{aligned} n_w(t) &= \mathrm{n}_1\varphi_1(t) + \mathrm{n}_2\varphi_2(t) + \cdots \\ &= \sum_k \mathrm{n}_k\varphi_k(t) \end{aligned}$$

White noise has infinite bandwidth. Consequently, the dimensionality of the signal space is infinity.

We shall now show that RVs $\mathrm{n}_1,\ \mathrm{n}_2,\ \ldots$ are independent, with variance $\mathcal{N}/2$ each. Because [see Eq. (14.13)]

$$\mathrm{n}_k = \int_{-\infty}^{\infty} \mathrm{n}_w(t)\varphi_k(t)\,dt$$

$$\overline{\mathrm{n}_j\mathrm{n}_k} = \overline{\int_{-\infty}^{\infty} \mathrm{n}_w(\alpha)\varphi_j(\alpha)\,d\alpha \int_{-\infty}^{\infty} \mathrm{n}_w(\beta)\varphi_k(\beta)\,d\beta}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overline{\mathrm{n}_w(\alpha)\mathrm{n}_w(\beta)}\varphi_j(\alpha)\varphi_k(\beta)\,d\alpha\,d\beta$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_{\mathrm{n}_w}(\beta-\alpha)\varphi_j(\alpha)\varphi_k(\beta)\,d\alpha\,d\beta$$

Because $R_{\mathrm{n}_w}(\tau) = \mathcal{F}^{-1}(\mathcal{N}/2) = (\mathcal{N}/2)\,\delta(\tau)$,

$$\overline{\mathrm{n}_j\mathrm{n}_k} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\mathcal{N}}{2}\delta(\beta-\alpha)\varphi_j(\alpha)\varphi_k(\beta)\,d\alpha\,d\beta$$

$$= \frac{\mathcal{N}}{2}\int_{-\infty}^{\infty}\varphi_j(\alpha)\varphi_k(\alpha)\,d\alpha$$

$$= \begin{cases} 0 & j \neq k \\ \dfrac{\mathcal{N}}{2} & j = k \end{cases} \tag{14.27}$$

Hence, n_j and n_k are uncorrelated gaussian RVs, each with variance $\mathcal{N}/2$. Since they are gaussian, uncorrelatedness implies independence. This proves the result.

For the time being, assume that we are considering an N-dimensional case. The joint PDF of independent joint gaussian RVs $\mathrm{n}_1, \mathrm{n}_2, \ldots, \mathrm{n}_N$, each with zero mean and variance $\mathcal{N}/2$, is [see Eq. (14.21)]

$$p_{\mathrm{n}}(\boldsymbol{n}) = \prod_{j=1}^{N}\frac{1}{\sqrt{2\pi\mathcal{N}/2}}e^{-n_j^2/2(\mathcal{N}/2)}$$

$$= \frac{1}{(\pi\mathcal{N})^{N/2}}e^{-(n_1^2+n_2^2+\cdots+n_N^2)/\mathcal{N}} \tag{14.28a}$$

$$= \frac{1}{(\pi\mathcal{N})^{N/2}}e^{-|\boldsymbol{n}|^2/\mathcal{N}} \tag{14.28b}$$

This shows that the PDF $p_{\mathbf{n}}(\boldsymbol{n})$ depends only on $|\boldsymbol{n}|$, the magnitude of the noise vector $\mathbf{n}$ in the hyperspace, and is therefore spherically symmetrical if plotted in the N-dimensional hyperspace.

14.3 OPTIMUM RECEIVER[2–5]

We shall now consider, from a more fundamental point of view, the problem of M-ary communication in the presence of **additive white gaussian channel noise (AWGN)**. No constraint shall be placed on the optimum structure, and we shall try to answer the fundamental question: What receiver will yield the minimum error probability?

The comprehension of the signal-detection problem is greatly facilitated by geometrical representation of signals. In a signal space, we can represent a signal by a fixed point (or a vector). A random process can be represented by a random point (or a random vector). The region in which the random point may lie will be shown shaded, with the shading intensity proportional to the probability of observing the signal in that region. In the M-ary scheme, we use M symbols, or messages, $m_1, m_2, \ldots, m_M$. Each of these symbols is represented by a

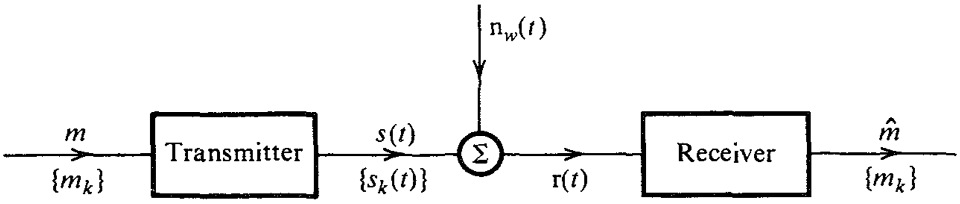


Figure 14.3 M-ary communication system.

specified waveform. Let the corresponding waveforms be $s_1(t)$, $s_2(t)$, ..., $s_M(t)$. Thus, the symbol (or message) m_k is sent by transmitting the waveform $s_k(t)$. These waveforms are corrupted by AWGN $n_w(t)$ (Fig. 14.3) with PSD

$$S_{n_w}(\omega) = \frac{\mathcal{N}}{2}$$

At the receiver, the received signal $r(t)$ consists of one of the M message waveforms $s_k(t)$ plus the channel noise,

$$r(t) = s_k(t) + n_w(t) \tag{14.29a}$$

We can represent $r(t)$ in a signal space by letting $\mathbf{r}$, s_k, and $\mathbf{n}_w$ be the points (or vectors) representing signals $r(t)$, $s_k(t)$, and $n_w(t)$, respectively, Then it is evident that

$$\mathbf{r} = s_k + \mathbf{n}_w \tag{14.29b}$$

The vector s_k is a fixed vector, because the waveform $s_k(t)$ is nonrandom. The vector $\mathbf{n}_w$ (or point $\mathbf{n}_w$) is random. Hence, the vector $\mathbf{r}$ is also random. Because $n_w(t)$ is a gaussian white noise, the probability distribution of $\mathbf{n}_w$ has spherical symmetry in the signal space (see Sec. 14.2). Hence, the distribution of $\mathbf{r}$ is a spherical distribution centered at a fixed point s_k, as shown in Fig. 14.4. Whenever the message m_k is transmitted, the probability of observing the received signal $r(t)$ in a given region is indicated by the intensity of the shading in Fig. 14.4. Actually, because the noise is white, the space has an infinite number of dimensions. For simplicity, however, we have shown the space to be 3-dimensional. This will suffice to indicate our line of reasoning. We can draw similar regions for various points s_1, s_2, ..., s_M. Figure 14.5a shows the regions for two messages m_j and m_k when s_j and s_k are widely separated in signal

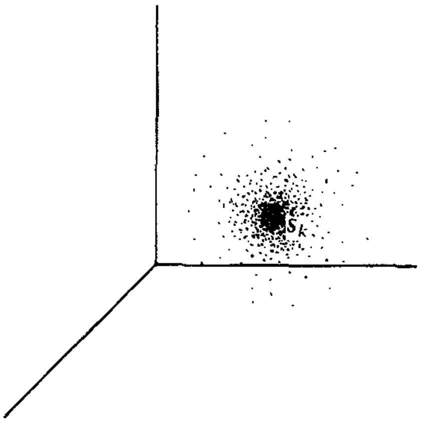


Figure 14.4 Effect of gaussian channel noise on the received signal.

space. In this case, there is virtually no overlap between the two regions. If either m_j or m_k is transmitted, the received signal will lie in one of the two regions. From the position of the received signal, one can decide with a very small probability of error whether m_j or m_k was transmitted. Note that theoretically each region extends to infinity, although the probability of observing the received signal diminishes rapidly as one moves away from the center. Hence, there will always be an overlap between the two regions, resulting in a nonzero error probability. In Fig. 14.5a, the received signal **r** is much closer to s_k than to s_j. It is therefore more likely that m_k was transmitted.

Figure 14.5b illustrates the case when the points s_j and s_k are spaced closely together. In this case, there is a considerable overlap between the two regions. Because the received signal r is closer to s_j than to s_k, it is more likely that m_j was transmitted. But in this case there is also a considerable probability that m_k may have been transmitted. Hence in this situation, there will be a much higher probability of error in any decision scheme.

The optimum receiver must decide which message has been transmitted from a knowledge of $\boldsymbol{r}$. The signal space must be divided into M nonoverlapping, or disjoint, regions

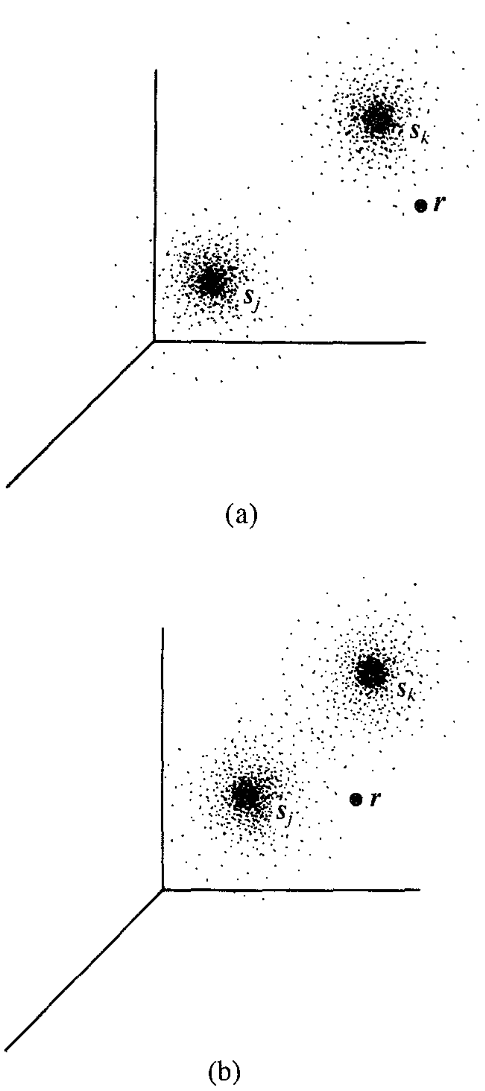


Figure 14.5 Binary communication in the presence of noise.

$R_1, R_2, \ldots, R_M$, corresponding to the M messages $m_1, m_2, \ldots, m_M$. If $\mathbf{r}$ falls in the region R_k, the decision is m_k. The problem of designing the receiver then reduces to choosing the boundaries of these regions $R_1, R_2, \ldots, R_M$ such that the probability of error in decision making is minimum.

To recapitulate: A transmitter transmits a sequence of messages from a set of M messages $m_1, m_2, \ldots, m_M$. These messages are represented by finite energy waveforms $s_1(t), s_2(t), \ldots, s_M(t)$. One waveform is transmitted every T_M seconds. We assume that the receiver is time synchronized with the transmitter. The waveforms are corrupted during transmissions by an AWGN of PSD $\mathcal{N}/2$. Knowing the received waveform, the receiver must make a decision as to which waveform was transmitted. The merit criterion of the receiver is the minimum probability of error in making this decision.

Let us now discuss the dimensionality of the signal space in our problem. If there was no noise, we would be dealing with only M waveforms $s_1(t), s_2(t), \ldots, s_M(t)$. In this case a signal space of, at most, M dimensions would suffice. This is because the dimensionality of a signal space is always equal to or less than the number of independent signals in the space (see Sec. 14.1). For the sake of generality we shall assume the space to have N dimensions ($N \leq M$). Let $\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t)$ be the orthonormal basis set for this space. Such a set can be constructed by using the Gram-Schmidt procedure discussed in Appendix C. We can then represent the signal waveform $s_k(t)$ as

$$\begin{aligned} s_k(t) &= s_{k_1}\varphi_1(t) + s_{k_2}\varphi_2(t) + \cdots + s_{kN}\varphi_N(t) \\ &= \sum_{j=1}^{N} s_{kj}\varphi_j(t) \end{aligned} \tag{14.30a}$$

where

$$s_{kj} = \int_{-\infty}^{\infty} s_k(t)\varphi_j(t)\,dt \tag{14.30b}$$

Now consider the white gaussian channel noise $\mathrm{n}_w(t)$. This signal has an infinite bandwidth ($B = \infty$). It has an infinite number of dimensions and obviously cannot be represented in the N-dimensional signal space discussed earlier. We can, however, split $\mathrm{n}_w(t)$ into two components: (1) the projection of $\mathrm{n}_w(t)$ on the N-dimensional signal space, and (2) the remaining component, which will be orthogonal to the N-dimensional signal space. Let us denote the two components by $\mathrm{n}(t)$ and $\mathrm{n}_0(t)$. Thus,

$$\mathrm{n}_w(t) = \mathrm{n}(t) + \mathrm{n}_0(t) \tag{14.31}$$

where

$$\mathrm{n}(t) = \sum_{k=1}^{N} \mathrm{n}_j\varphi_j(t) \tag{14.32a}$$

and

$$\mathrm{n}_j = \int_{-\infty}^{\infty} \mathrm{n}(t)\varphi_j(t)\,dt \tag{14.32b}$$

Because $\mathrm{n}_0(t)$ is orthogonal to the N-dimensional space, it is orthogonal to every signal in that space. Hence,

$$\int_{-\infty}^{\infty} \mathrm{n}_0(t)\varphi_j(t)\,dt = 0 \qquad j = 1, 2, \ldots, N$$

Therefore,

$$\begin{aligned}\mathrm{n}_j &= \int_{-\infty}^{\infty} [\mathrm{n}(t) + \mathrm{n}_0(t)]\varphi_j(t)\,dt \\ &= \int_{-\infty}^{\infty} \mathrm{n}_w(t)\varphi_j(t)\,dt \end{aligned} \tag{14.33}$$

From Eqs. (14.33) and (14.32a) it is evident that we can filter out the component $\mathrm{n}_0(t)$ from $\mathrm{n}_w(t)$. This can be seen from the fact that the received signal, $\mathrm{r}(t)$, can be expressed as

$$\begin{aligned}\mathrm{r}(t) &= s_k(t) + \mathrm{n}_w(t) \\ &= s_k(t) + \mathrm{n}(t) + \mathrm{n}_0(t) \\ &= \mathrm{q}(t) + \mathrm{n}_0(t)\end{aligned} \tag{14.34}$$

where $\mathrm{q}(t)$ is the projection of $\mathrm{r}(t)$ on the N-dimensional space. Thus,

$$\mathrm{q}(t) = s_k(t) + \mathrm{n}(t) \tag{14.35}$$

We can obtain the projection $\mathrm{q}(t)$ from $\mathrm{r}(t)$ by observing that [see Eqs. (14.30) and (14.32a)]

$$\mathrm{q}(t) = \sum_{j=1}^{M} (s_{kj} + \mathrm{n}_j)\varphi_j(t) \tag{14.36}$$

From Eqs. (14.30b), (14.33), and (14.36) it follows that if we feed the received signal $\mathrm{r}(t)$ into the system shown in Fig. 14.6, the resultant outcome will be $\mathrm{q}(t)$. Thus, the orthogonal noise component can be filtered out without disturbing the message signal. The question here is: Would such filtering help in our decision making? We can easily show that it cannot hurt us. The noise $\mathrm{n}_w(t)$ is independent of the signal waveform $s_k(t)$. Therefore, its component $\mathrm{n}_0(t)$ is also independent of $s_k(t)$. Thus, $\mathrm{n}_0(t)$ contains no information about the transmitted signal, and discarding such a component from the received signal $\mathrm{r}(t)$ will not cause any loss of information regarding the signal waveform $s_k(t)$. This, however, is not enough. We must also make sure that the noise being discarded $[\mathrm{n}_0(t)]$ is not in any way related to the remaining noise component $\mathrm{n}(t)$. If $\mathrm{n}_0(t)$ and $\mathrm{n}(t)$ are related in any way, it will be possible to obtain some information about $\mathrm{n}(t)$ from $\mathrm{n}_0(t)$, thus enabling us to detect that signal with less error probability. If the components $\mathrm{n}_0(t)$ and $\mathrm{n}(t)$ are independent random processes, the component $\mathrm{n}_0(t)$ does not carry any information about $\mathrm{n}(t)$ and can be discarded. Under these conditions, $\mathrm{n}_0(t)$ is **irrelevant** to the decision making at the receiver.

The process $\mathrm{n}(t)$ is represented by components n_1, n_2, $\ldots$, n_N along $\varphi_1(t)$, $\varphi_2(t)$, $\ldots$, $\varphi_N(t)$, and $\mathrm{n}_0(t)$ is represented by the remaining components (infinite number) along the remaining basis signals in the complete set, $\{\varphi_k(t)\}$. From Eq. (14.27) we observe that all the components are independent. Hence, the components representing $\mathrm{n}_0(t)$ are independent of the components representing $\mathrm{n}(t)$. Consequently, $\mathrm{n}_0(t)$ is independent of $\mathrm{n}(t)$ and is irrelevant data.

The received signal $\mathrm{r}(t)$ is now reduced to the signal $\mathrm{q}(t)$, which contains the desired signal waveform and the projection of the channel noise on the N-dimensional signal space.

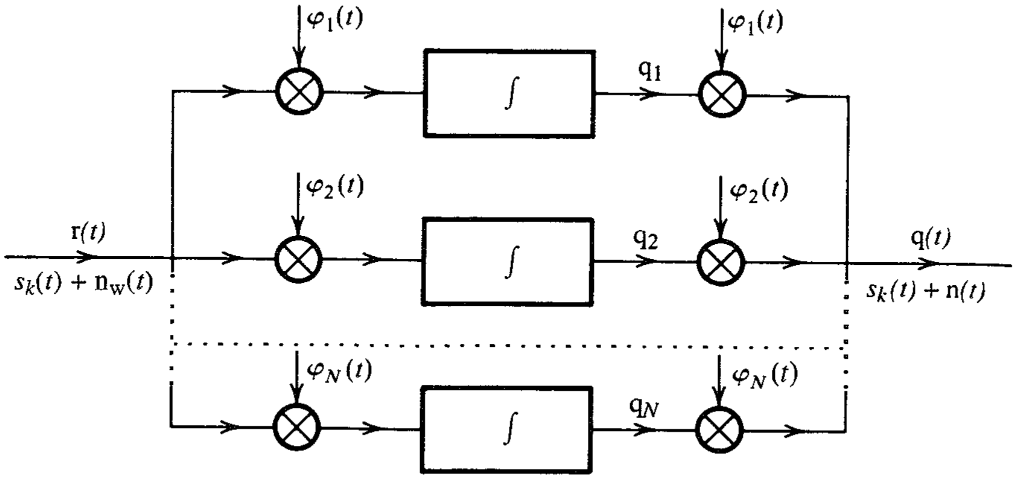


Figure 14.6 Eliminating the noise orthogonal to signal space.

Thus, the signal $q(t)$ can be completely represented in the signal space. Let the vectors representing $n(t)$ and $q(t)$ be denoted by $\mathbf{n}$ and $\mathbf{q}$. Thus,

$$\mathbf{q} = s + \mathbf{n}$$

where s may be any one of vectors $s_1, s_2, \ldots, s_M$.

The random vector $\mathbf{n}$ $(n_1, n_2, \ldots, n_N)$ is represented by N independent Gaussian variables, each with zero mean and variance $\sigma_n{}^2 = \mathcal{N}/2$. The joint PDF of vector $\mathbf{n}$ in such a case has a spherical symmetry, as shown in Eq. (14.28b),

$$p_{\mathbf{n}}(\boldsymbol{n}) = \frac{1}{(\pi\mathcal{N})^{N/2}} e^{-|\boldsymbol{n}|^2/\mathcal{N}} \tag{14.37a}$$

Note that this is actually a compact notation for

$$p_{n_1,n_2,\ldots,n_N}(n_1, n_2, \ldots, n_N) = \frac{1}{(\pi\mathcal{N})^{N/2}} e^{-(n_1^2+n_2^2+\cdots+n_N^2)/\mathcal{N}} \tag{14.37b}$$

Decision Procedure

Our problem is now considerably simplified. The irrelevant noise component has been filtered out. The residual signal $q(t)$ can be represented in an N-dimensional signal space. We proceed to determine the M decision regions $R_1, R_2, \ldots, R_M$ in this space. The regions must be so chosen that the probability of error in making the decision is minimized.

Suppose the received vector $\mathbf{q} = \boldsymbol{q}$. Then if the receiver decides $\hat{m} = m_k$, the conditional probability of making the correct decision, given that $\mathbf{q} = \boldsymbol{q}$, is

$$P(C|\mathbf{q} = \boldsymbol{q}) = P(m_k|\mathbf{q} = \boldsymbol{q}) \tag{14.38}$$

where $P(C|\mathbf{q} = \boldsymbol{q})$ is the conditional probability of making the correct decision given $\mathbf{q} = \boldsymbol{q}$, and $P(m_k|\mathbf{q} = \boldsymbol{q})$ is the conditional probability that m_k was transmitted given $\mathbf{q} = \boldsymbol{q}$. The unconditional probability $P(C)$ is given by

$$P(C) = \int_{\boldsymbol{q}} P(C|\mathbf{q} = \boldsymbol{q}) p_{\mathbf{q}}(\boldsymbol{q})\, d\boldsymbol{q} \tag{14.39}$$

where the integration is performed over the entire region occupied by $\mathbf{q}$. Note that this is an N-fold integration with respect to the variables $q_1, q_2, \ldots, q_N$ over the range $(-\infty, \infty)$. Also, because $p_\mathbf{q}(\boldsymbol{q}) \geq 0$, this integral is maximum when $P(C|\mathbf{q} = \boldsymbol{q})$ is maximum. From Eq. (14.38) it now follows that if a decision $\hat{m} = m_k$ is made, the error probability is minimized if

$$P(m_k|\mathbf{q} = \boldsymbol{q})$$

is maximized. The probability $P(m_k|\mathbf{q} = \boldsymbol{q})$ is called the **a posteriori probability** of m_k. This is because it represents the probability that m_k was transmitted when $\boldsymbol{q}$ is received.

The decision procedure is now clear. Once we receive $\mathbf{q} = \boldsymbol{q}$, we evaluate all M a posteriori probabilities. Then we make the decision in favor of that message for which the a posteriori probability is highest—that is, the receiver decides that $\hat{m} = m_k$ if

$$P(m_k|\mathbf{q} = \boldsymbol{q}) > P(m_j|\mathbf{q} = \boldsymbol{q}) \qquad \text{for all } j \neq k \tag{14.40}$$

Thus, the detector that minimizes the error probability is the **maximum a posteriori probability (MAP) detector.**

We can use Bayes' mixed rule [Eq. (10.49b)] to determine the a posteriori probabilities. We have

$$P(m_k|\mathbf{q} = \boldsymbol{q}) = \frac{P(m_k)p_\mathbf{q}(\boldsymbol{q}|m_k)}{p_\mathbf{q}(\boldsymbol{q})} \tag{14.41}$$

Hence, the receiver decides $\hat{m} = m_k$ if the decision function

$$\frac{P(m_i)p_\mathbf{q}(\boldsymbol{q}|m_i)}{p_\mathbf{q}(\boldsymbol{q})} \qquad i = 1, 2, \ldots, M$$

is maximum for $i = k$.

Note that the denominator $p_\mathbf{q}(\boldsymbol{q})$ is common to all decision functions and, hence, may be ignored. Thus, the receiver sets $\hat{m} = m_k$ if the decision function

$$P(m_i)p_\mathbf{q}(\boldsymbol{q}|m_i) \qquad i = 1, 2, \ldots, M \tag{14.42}$$

is maximum for $i = k$.

Thus, once $\boldsymbol{q}$ is obtained, we compute the decision function [Eq. (14.42)] for all messages $m_1, m_2, \ldots, m_M$ and decide that the message for which the function is maximum is the one most likely to have been sent.

We now turn our attention to computing the decision functions. The a priori probability $P(m_i)$ represents the probability that the message m_i will be transmitted. These probabilities must be known if the criterion discussed is to be used.* The term $p_\mathbf{q}(\boldsymbol{q}|m_i)$ represents the PDF of $\mathbf{q}$ when $s(t) = s_i(t)$. Under this condition,

$$\mathbf{q} = s_i + \mathbf{n}$$

and

$$\mathbf{n} = \mathbf{q} - s_i$$

The point s_i is constant, and $\mathbf{n}$ is a random point. Obviously, $\mathbf{q}$ is a random point with the same distribution as $\mathbf{n}$ but centered at the points s_i.

---

* In case these probabilities are unknown, one must use other merit criteria, such as maximum likelihood or minimax, as will be discussed in later sections.

Alternately, the probability $\mathbf{q} = \boldsymbol{q}$ (given $\mathrm{m} = m_i$) is the same as the probability $\mathbf{n} = \boldsymbol{q} - \boldsymbol{s}_i$. Hence [Eq. (14.37a)],

$$p_{\mathbf{q}}(\boldsymbol{q}|m_i) = p_{\mathbf{n}}(\boldsymbol{q} - \boldsymbol{s}_i) = \frac{1}{(\pi \mathcal{N})^{N/2}} e^{-|\boldsymbol{q}-\boldsymbol{s}_i|^2/\mathcal{N}} \tag{14.43}$$

The decision function in Eq. (14.42) now becomes

$$\frac{P(m_i)}{(\pi \mathcal{N})^{N/2}} e^{-|\boldsymbol{q}-\boldsymbol{s}_i|^2/\mathcal{N}} \tag{14.44}$$

Note that the decision function is always nonnegative for all values of i. Hence, comparing these functions is equivalent to comparing their logarithms, because the logarithm is a monotone function for the positive argument. Hence, for convenience, the decision function will be chosen as the logarithm of Eq. (14.44). In addition, the factor $(\pi \mathcal{N})^{N/2}$ is common for all i and can be left out. Hence, the decision function is

$$\ln P(m_i) - \frac{1}{\mathcal{N}} |\boldsymbol{q} - \boldsymbol{s}_i|^2 \tag{14.45}$$

Note that $|\boldsymbol{q} - \boldsymbol{s}_i|^2$ is the square of the length of the vector $\boldsymbol{q} - \boldsymbol{s}_i$. Hence,

$$\begin{aligned} |\boldsymbol{q} - \boldsymbol{s}_i|^2 &= (\boldsymbol{q} - \boldsymbol{s}_i) \cdot (\boldsymbol{q} - \boldsymbol{s}_i) \\ &= |\boldsymbol{q}|^2 + |\boldsymbol{s}_i|^2 - 2\boldsymbol{q} \cdot \boldsymbol{s}_i \end{aligned} \tag{14.46}$$

Hence, the decision function in Eq. (14.45) becomes (after multiplying throughout by $\mathcal{N}/2$)

$$\frac{\mathcal{N}}{2} \ln P(m_i) - \frac{1}{2}[|\boldsymbol{q}|^2 + |\boldsymbol{s}_i|^2 - 2\boldsymbol{q} \cdot \boldsymbol{s}_i] \tag{14.47}$$

Note that the term $|\boldsymbol{s}_i|^2$ is the square of the length of $\boldsymbol{s}_i$ and represents E_i, the energy of signal $s_i(t)$. The terms $\mathcal{N} \ln P(m_i)$ and E_i are constants in the decision function. Let

$$a_i = \tfrac{1}{2}[\mathcal{N} \ln P(m_i) - E_i] \tag{14.48}$$

Now the decision function in Eq. (14.47) becomes

$$a_i + \boldsymbol{q} \cdot \boldsymbol{s}_i - \tfrac{1}{2}|\boldsymbol{q}|^2$$

The term $\frac{1}{2}|\boldsymbol{q}|^2$ is common to all M decision functions and can be omitted for the purpose of comparison. Thus, the new decision function b_i is

$$b_i = a_i + \boldsymbol{q} \cdot \boldsymbol{s}_i \tag{14.49}$$

We compute this function b_i for $i = 1, 2, \ldots, N$, and the receiver decides that $\hat{m} = m_k$ if this function is the largest for $i = k$. If the signal $q(t)$ is applied at the input terminals of a system whose impulse response is $h(t)$, the output at $t = T_M$ is given by

$$\int_{-\infty}^{\infty} q(\lambda) h(T_M - \lambda)\, d\lambda$$

If we choose a filter matched to $s_i(t)$, that is, $h(t) = s_i(T_M - t)$,

$$h(T_M - \lambda) = s_i(\lambda)$$

and the output is

$$\int_{-\infty}^{\infty} q(\lambda)s_i(\lambda)\,d\lambda = \boldsymbol{q} \cdot \boldsymbol{s}_i$$

Hence, $\boldsymbol{q} \cdot \boldsymbol{s}_i$ is the output at $t = T_M$ of a filter matched to $s_i(t)$ when $q(t)$ is applied to its input.

Actually, we do not have $q(t)$. The incoming signal $r(t)$ is given by

$$\begin{aligned} r(t) &= s_i(t) + n_w(t) \\ &= \underbrace{s_i(t) + n(t)}_{q(t)} + \underbrace{n_0(t)}_{\substack{\text{irrelevant} \\ \text{noise}}} \end{aligned}$$

where $n_0(t)$ is the (irrelevant) component of $n_w(t)$ orthogonal to the N-dimensional signal space. Because $n_0(t)$ is orthogonal to this space, it is orthogonal to every signal in this space. Hence, it is orthogonal to the signal $s_i(t)$, and

$$\int_{-\infty}^{\infty} n_0(t)s_i(t)\,dt = 0$$

and

$$\begin{aligned} \boldsymbol{q} \cdot \boldsymbol{s}_i &= \int_{-\infty}^{\infty} q(t)s_i(t)\,dt + \int_{-\infty}^{\infty} n_0(t)s_i(t)\,dt \\ &= \int_{-\infty}^{\infty} [q(t) + n_0(t)]s_i(t)\,dt \\ &= \int_{-\infty}^{\infty} r(t)s_i(t)\,dt \end{aligned} \tag{14.50}$$

Hence, it is immaterial whether we use $q(t)$ or $r(t)$ at the input. We thus apply the incoming signal $r(t)$ to a parallel bank of matched filters, and the output of the filters is sampled at $t = T_M$. Then a constant a_i is added to the ith filter output sample, and the resulting outputs are compared. The decision is made in favor of the signal for which this output is the largest. The receiver implementation for this decision procedure is shown in Fig. 14.7a. As shown in Chapter 13, a matched filter is equivalent to a correlator. One may therefore use correlators instead of matched filters. Such an arrangement is shown in Fig. 14.7b.

We have shown that in the presence of AWGN, the matched-filter receiver is the optimum receiver when the merit criterion is the minimum error probability. Note that the system is linear, although it was not constrained to be so. Therefore, for white gaussian noise the optimum receiver happens to be linear. The matched filter obtained in Chapter 13 and the decision procedure are identical to those derived here.

The optimum receiver can be implemented in another way. From Eq. (14.50), we have

$$\boldsymbol{q} \cdot \boldsymbol{s}_i = \boldsymbol{r} \cdot \boldsymbol{s}_i$$

From Eq. (14.3), we can rewrite this as

$$\boldsymbol{q} \cdot \boldsymbol{s}_i = \sum_{j=1}^{N} r_j s_{ij}$$

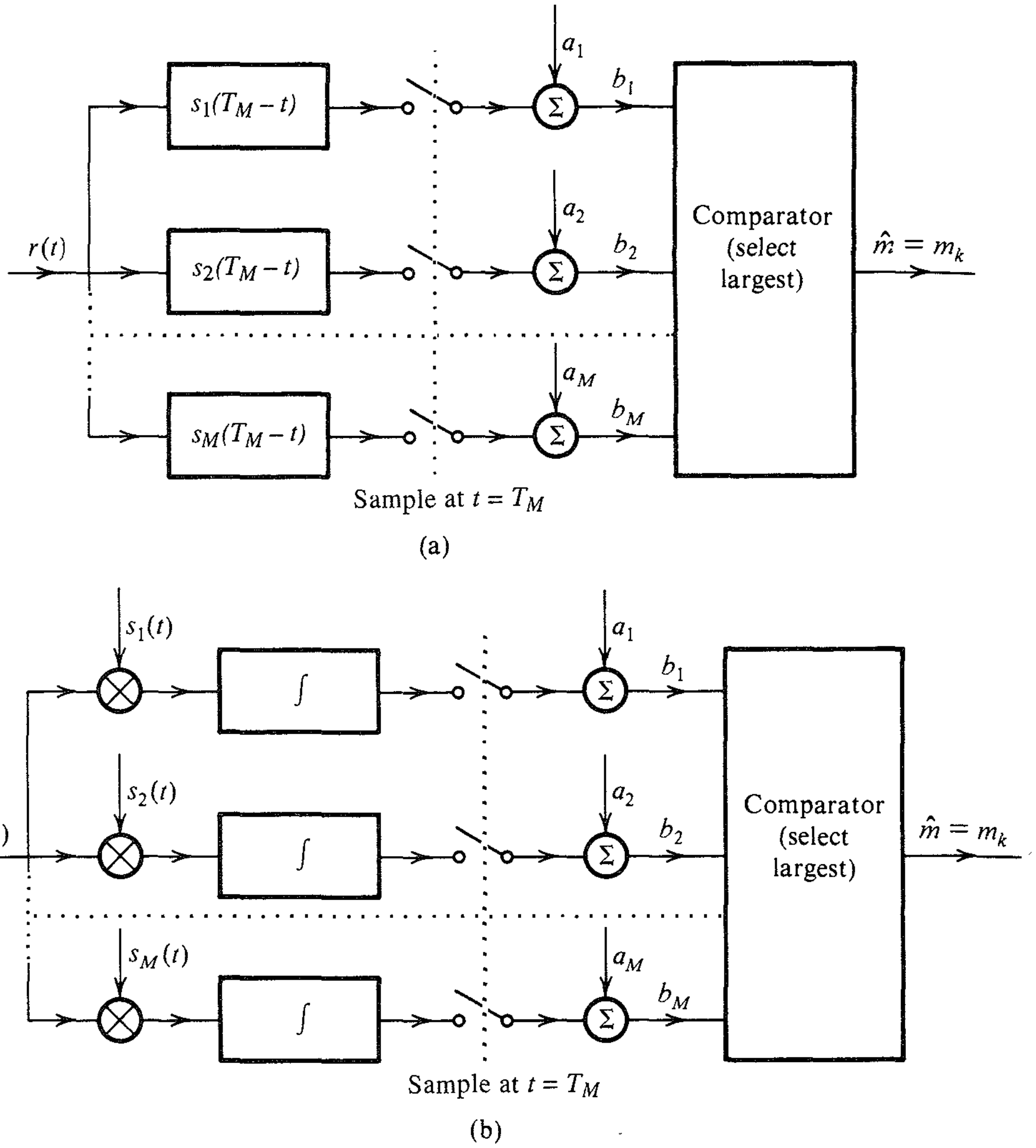

$r(t)$
$s_1(T_M - t)$
$s_2(T_M - t)$
$s_M(T_M - t)$
a_1
a_2
a_M
b_1
b_2
b_M
Σ
Comparator (select largest)
$\hat{m} = m_k$
Sample at $t = T_M$
(a)
$r(t)$
$s_1(t)$
$s_2(t)$
$s_M(t)$
∫
a_1
a_2
a_M
b_1
b_2
b_M
Σ
Comparator (select largest)
$\hat{m} = m_k$
Sample at $t = T_M$
(b)

Figure 14.7 Optimum M-ary receiver. (a) Matched-filter detector. (b) Correlation detector.

The term $q \cdot s_i$ is computed according to this equation by first generating r_j's and then computing the sum of $r_j s_{ij}$ (remember that the s_{ij}'s are known), as shown in Fig. 14.8a. The M correlator detectors in Fig. 14.7b can be replaced by N filters matched to $\varphi_1(t)$, $\varphi_2(t)$, …, $\varphi_N(t)$, as shown in Fig. 14.8b. Both types of optimum receivers (Figs. 14.7 and 14.8) perform identically. The choice will depend on the circumstances. For example, if $N < M$ and signals $\{\varphi_j(t)\}$ are easier to generate than $\{s_j(t)\}$, then the choice of Fig. 14.8 is obvious.

Decision Regions and Error Probability

In order to compute the error probability of the optimum receiver, we need to determine decision regions in the signal space first. As mentioned earlier, the signal space is divided into M nonoverlapping, or disjoint, decision regions R_1, R_2, …, R_M, corresponding to M

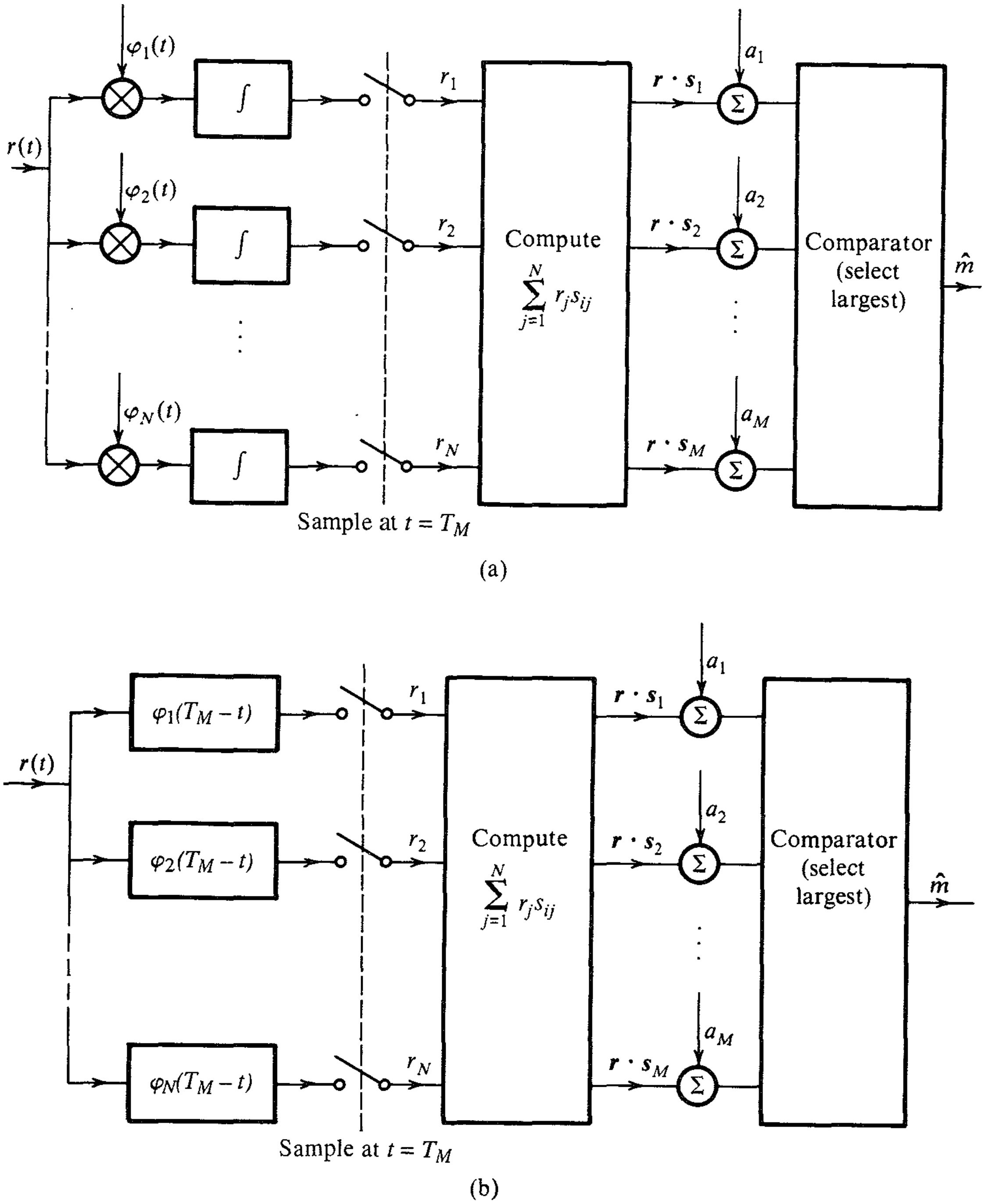

$\varphi_1(t)$
$\varphi_2(t)$
$\varphi_N(t)$
$r(t)$
$\int$
r_1
r_2
r_N
Compute $\sum_{j=1}^{N} r_j s_{ij}$
$r \cdot s_1$
$r \cdot s_2$
$r \cdot s_M$
a_1
a_2
a_M
Σ
Comparator (select largest)
$\hat{m}$
Sample at $t = T_M$
(a)
$\varphi_1(T_M - t)$
$\varphi_2(T_M - t)$
$\varphi_N(T_M - t)$
(b)

Figure 14.8 Another form of optimum M-ary receiver. (a) Correlation detector. (b) Matched-filter detector.

messages. If q falls in the region R_k, the decision is that m_k was transmitted. The decision regions are so chosen that the probability of error of the receiver is minimum. In light of this geometrical representation, we shall now try to interpret how the optimum receiver sets these decision regions.

The decision function is given by Eq. (14.45). The optimum receiver sets $\hat{m} = m_k$ if the decision function

$$\mathcal{N} \ln P(m_i) - |q - s_i|^2$$

is maximum for $i = k$. This equation defines the decision regions.

For simplicity, let us first consider the case of equiprobable messages; that is, $P(m_i) = 1/M$ for all i. In this case, the first term in the decision function is the same for all i and, hence, can be dropped. Thus, the receiver decides that $\hat{m} = m_k$ if the term $-|q - s_i|^2$ is largest (numerically the smallest) for $i = k$. Alternatively, this may also be stated as follows: the receiver decides that $\hat{m} = m_k$ if the decision function $|q - s_i|^2$ is minimum for $i = k$. Note that $|q - s_i|$ is the distance of point q from point s_i. Thus, the decision procedure in this case has a simple interpretation in geometrical space. The decision is made in favor of that signal which is closest to q, the projection of r [the component of $r(t)$] in the signal space. This result is expected on qualitative grounds for gaussian noise, because the gaussian noise has a spherical symmetry. If, however, the messages are not equiprobable, we cannot go too far on purely qualitative grounds. Nevertheless, we can draw certain broad conclusions. If a particular message m_i is more likely than the others, one will be safer in deciding more often in favor of m_i than other messages. Hence, in such a case the decision regions will be biased, or weighted, in favor of m_i. This is shown by the appearance of the term $\ln P(m_i)$ in the decision function. To better understand this point, let us consider a 2-dimensional signal space and two signals s_1 and s_2, as shown in Fig. 14.9a. In this figure, the decision regions R_1 and R_2 are shown for equiprobable messages; $P(m_1) = P(m_2) = 0.5$. The boundary of the decision region is the perpendicular bisector of the line joining points s_1 and s_2. Note that any point on the boundary is equidistant from s_1 and s_2. If q happens to fall on the boundary, we just "flip a coin" and decide whether to select m_1 or m_2. Figure 14.9b shows the case when the two messages are not equiprobable. To delineate the boundary of the decision regions, we use Eq. (14.45). The decision is m_1 if

$$|q - s_1|^2 - \mathcal{N} \ln P(m_1) < |q - s_2|^2 - \mathcal{N} \ln P(m_2)$$

Otherwise, the decision is m_2.

Note that $|q - s_1|$ and $|q - s_2|$ represent d_1 and d_2, the distance of q from s_1 and s_2, respectively. Thus, the decision is m_1 if

$$d_1^2 - d_2^2 < \mathcal{N} \ln \frac{P(m_1)}{P(m_2)}$$

The right-hand side of this inequality is a constant c:

$$c = \mathcal{N} \ln \frac{P(m_1)}{P(m_2)}$$

Thus, the decision is m_1 if

$$d_1^2 - d_2^2 < c$$

The decision is m_2 if

$$d_1^2 - d_2^2 > c$$

On the boundary of the decision regions,

$$d_1^2 - d_2^2 = c$$

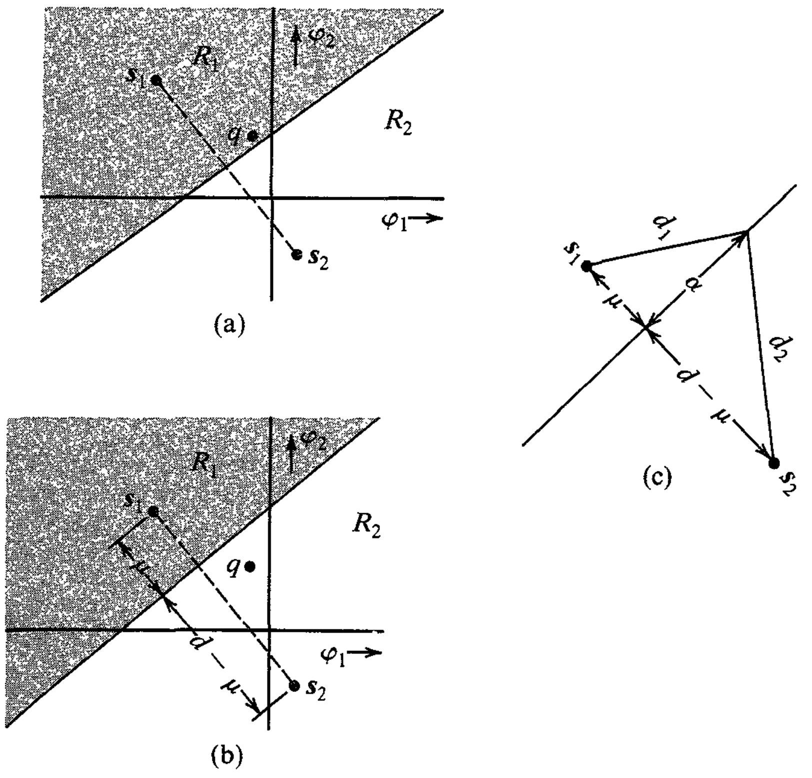


Figure 14.9 Determining optimum decision regions in a binary case.

We now show that such a boundary is given by a straight line perpendicular to line $s_1 s_2$ and passing through $s_1 s_2$ at a distance μ from s_1, where

$$\mu = \frac{c + d^2}{2d} = \frac{\mathcal{N}}{2d} \ln \left[\frac{P(m_1)}{P(m_2)} \right] + \frac{d}{2} \tag{14.51}$$

where d is the distance between s_1 and s_2. To prove this, we redraw the pertinent part of Fig. 14.9b as Fig. 14.9c. It is evident from this figure that

$$d_1^2 = \alpha^2 + \mu^2$$

$$d_2^2 = \alpha^2 + (d - \mu)^2$$

Hence,

$$d_1^2 - d_2^2 = 2d\mu - d^2 = c$$

Therefore,

$$\mu = \frac{c + d^2}{2d}$$

This is the desired result. Thus, along the decision boundary $d_1^2 - d_2^2$ is constant and equal to c. The boundaries of the decision regions for $M > 2$ may be determined along similar lines. The decision regions for the case of three equiprobable 2-dimensional signals are shown in Fig. 14.10. The boundaries of the decision regions are perpendicular bisectors of the lines

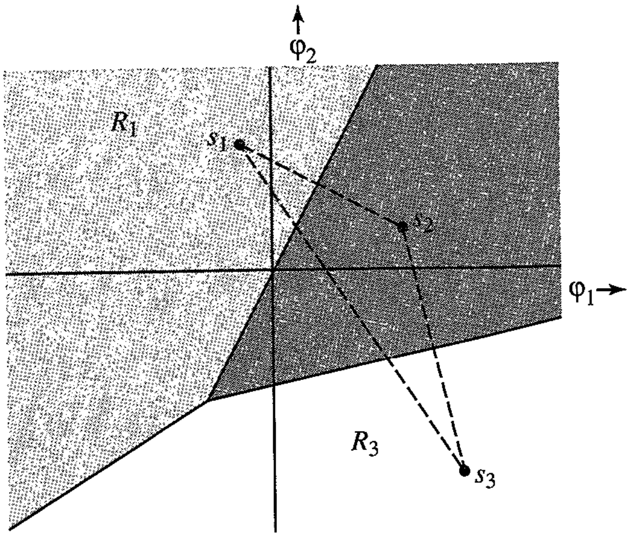


Figure 14.10 Determining optimum decision regions.

joining the original transmitted signals. If the signals are not equiprobable, then the boundaries will be shifted away from the signals with larger probabilities of occurrence.

For signals in N-dimensional space, the decision regions will be N-dimensional hypercones.

If there are M messages $m_1, m_2, \ldots, m_M$ with decision regions $R_1, R_2, \ldots, R_M$, respectively, then $P(C|m_i)$, the probability of a correct decision when m_i is transmitted, is given by

$$P(C|m_i) = P(q \text{ lies in } R_1) \tag{14.52}$$

and $P(C)$, the probability of a correct decision, is given by

$$P(C) = \sum_{i=1}^{M} P(m_i)P(C|m_i) \tag{14.53a}$$

and P_{eM}, the probability of error, is given by

$$P_{eM} = 1 - P(C) \tag{14.53b}$$

EXAMPLE 14.2 Binary data is transmitted using polar signaling over an AWGN channel with noise PSD $\mathcal{N}/2$. The two signals used are

$$s_1(t) = -s_2(t) = \sqrt{E}\varphi(t)$$

The symbol probabilities $P(m_1)$ and $P(m_2)$ are unequal. Design the optimum receiver and determine the corresponding error probability.

The two signals are represented graphically in Fig. 14.11a. If the energy of each signal is E, the distance of each signal from the origin is $\sqrt{E}$. The distance d between the two signals is

$$d = 2\sqrt{E}$$

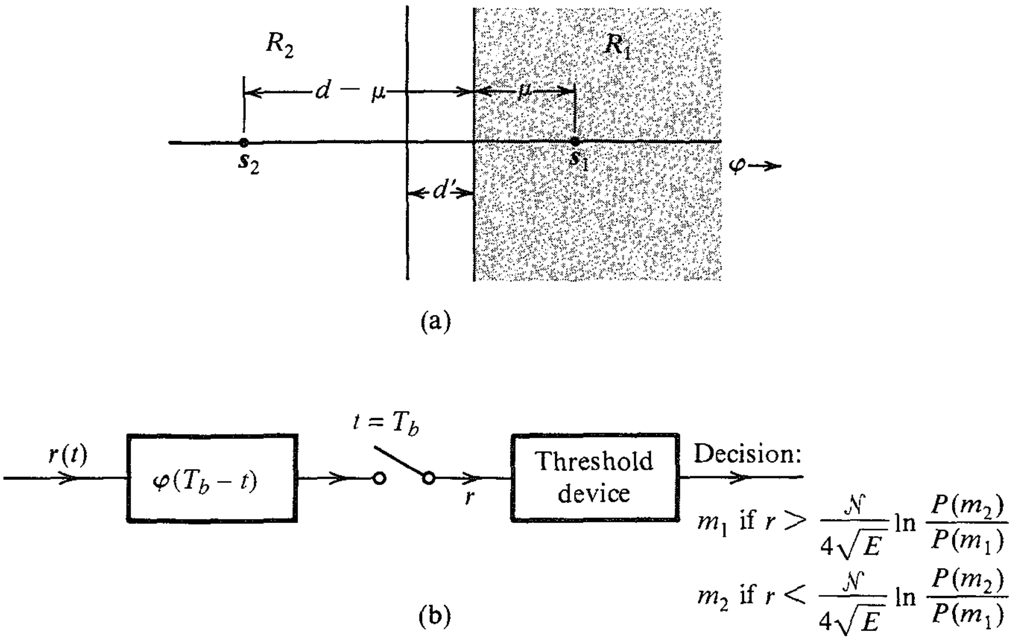

R_2
$d-\mu$
μ
R_1
s_2
s_1
d'
$\varphi\rightarrow$
(a)
$r(t)$
$\varphi(T_b-t)$
$t=T_b$
r
Threshold device
Decision:
m_1 if $r>\frac{\mathcal{N}}{4\sqrt{E}}\ln\frac{P(m_2)}{P(m_1)}$
m_2 if $r<\frac{\mathcal{N}}{4\sqrt{E}}\ln\frac{P(m_2)}{P(m_1)}$
(b)

Figure 14.11 Decision regions for a binary case in Example 14.2.

The decision regions R_1 and R_2 are shown in Fig. 14.11a. The distance μ is given by Eq. (14.51). Also,

$$\begin{aligned}P(C|\mathrm{m}=m_1) &= P(\text{noise vector originating at } s_1 \text{ remains in } R_1)\\ &= P(n>-\mu)\\ &= 1-Q\left(\frac{\mu}{\sigma_n}\right)\\ &= 1-Q\left(\frac{\mu}{\sqrt{\mathcal{N}/2}}\right)\end{aligned}$$

Similarly,

$$P(C|\mathrm{m}=m_2)=1-Q\left(\frac{d-\mu}{\sqrt{\mathcal{N}/2}}\right)$$

and

$$\begin{aligned}P(C) &= P(m_1)\left[1-Q\left(\frac{\mu}{\sqrt{\mathcal{N}/2}}\right)\right]+P(m_2)\left[1-Q\left(\frac{d-\mu}{\sqrt{\mathcal{N}/2}}\right)\right]\\ &= 1-P(m_1)Q\left(\frac{\mu}{\sqrt{\mathcal{N}/2}}\right)-P(m_2)Q\left(\frac{d-\mu}{\sqrt{\mathcal{N}/2}}\right)\end{aligned}$$

and

$$P_e=1-P(C)=P(m_1)Q\left(\frac{\mu}{\sqrt{\mathcal{N}/2}}\right)+P(m_2)Q\left(\frac{d-\mu}{\sqrt{\mathcal{N}/2}}\right) \tag{14.54a}$$

where

$$d = 2\sqrt{E} \tag{14.54b}$$

and

$$\mu = \frac{\mathcal{N}}{4\sqrt{E}} \ln \frac{P(m_1)}{P(m_2)} + \sqrt{E} \tag{14.54c}$$

When $P(m_1) = P(m_2) = 0.5$, $\mu = \sqrt{E} = d/2$, and Eq. (14.54a) reduces to

$$P_e = Q\left(\sqrt{\frac{2E}{\mathcal{N}}}\right) \tag{14.54d}$$

In this problem, because $N = 1$ and $M = 2$, the receiver in Fig. 14.8 is preferable to that in Fig. 14.7. For this case the receiver of the form in Fig. 14.8b reduces to that shown in Fig. 14.11b. The decision threshold d' as seen from Fig. 14.11a is

$$d' = \sqrt{E} - \mu = \frac{\mathcal{N}}{4\sqrt{E}} \ln \frac{P(m_2)}{P(m_1)}$$

Note that d' is the decision threshold. Thus, in Fig. 14.11b, if the receiver output $r > d'$, the decision is m_1. Otherwise the decision is m_2.

When $P(m_1) = P(m_2) = 0.5$, the decision threshold is zero. This is precisely the result derived in Chapter 13 for polar signaling.

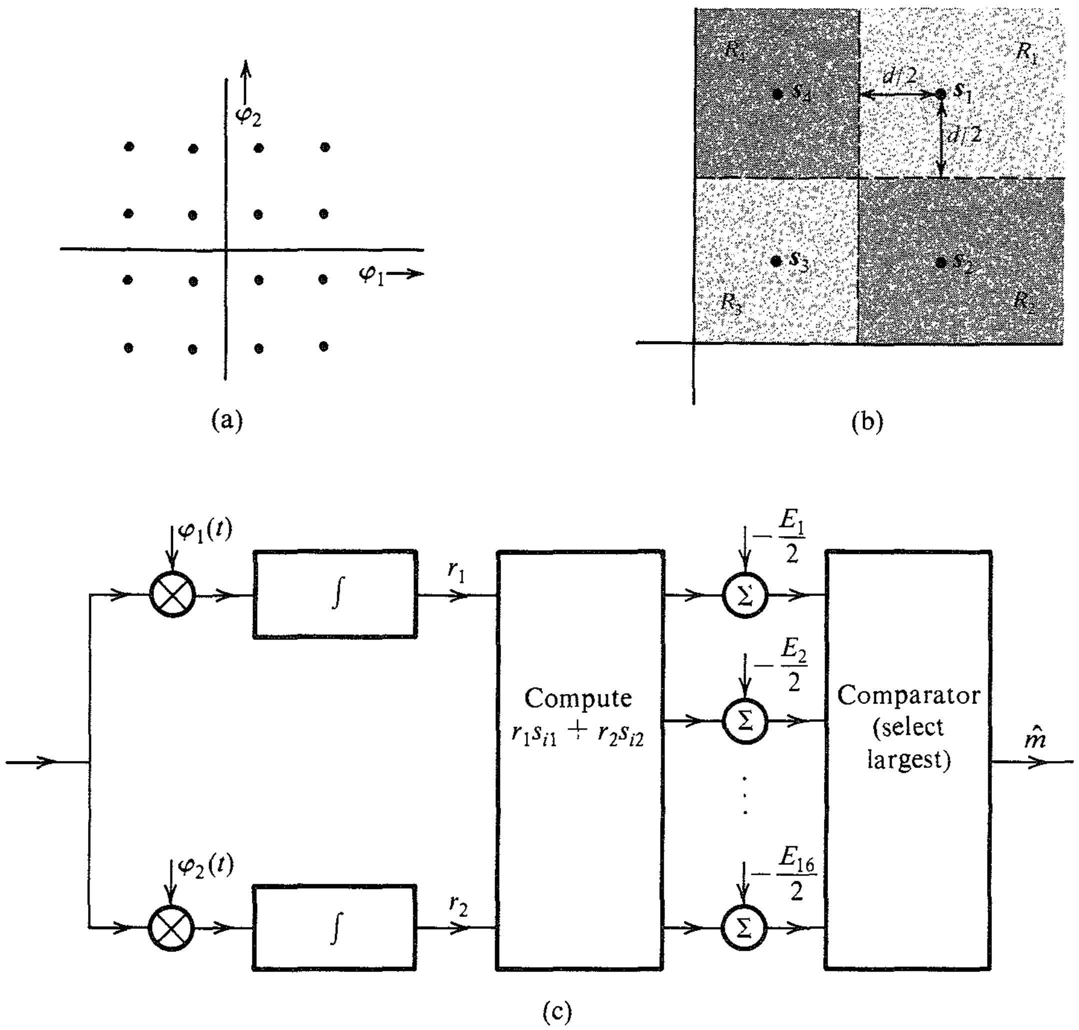
EXAMPLE 14.3 QAM

Design the optimum receiver and compute the corresponding error probability for the 16-point QAM configuration shown in Fig. 14.12a, assuming all signals to be equiprobable and assuming an AWGN channel.

Let us first calculate the error probability. The first quadrant of the signal space is reproduced in Fig. 14.12b. Because all the signals are equiprobable, the decision region boundaries will be perpendicular bisectors joining various signals, as shown in Fig. 14.12b.

From Fig. 14.12b it follows that

$$\begin{aligned} P(C|m_1) &= P(\text{noise vector originating at } s_1 \text{ lies within } R_1) \\ &= P\left(\mathrm{n}_1 > -\frac{d}{2},\ \mathrm{n}_2 > -\frac{d}{2}\right) \\ &= P\left(\mathrm{n}_1 > -\frac{d}{2}\right) P\left(\mathrm{n}_2 > -\frac{d}{2}\right) \\ &= \left[1 - Q\left(\frac{d/2}{\sigma_{\mathrm{n}}}\right)\right]^2 \\ &= \left[1 - Q\left(\frac{d}{\sqrt{2\mathcal{N}}}\right)\right]^2 \end{aligned}$$


φ_2
φ_1
R_4
s_4
R_1
$d/2$
s_1
$d/2$
s_3
R_3
s_2
R_2
$\varphi_1(t)$
$\int$
r_1
$\varphi_2(t)$
$\int$
r_2
Compute $r_1 s_{i1} + r_2 s_{i2}$
$-\frac{E_1}{2}$
Σ
$-\frac{E_2}{2}$
Σ
$-\frac{E_{16}}{2}$
Σ
Comparator (select largest)
$\hat{m}$

(a) (b)

(c)

Figure 14.12 16-ary QAM.

For convenience, let us define

$$p = 1 - Q\left(\frac{d}{\sqrt{2\mathcal{N}}}\right) \tag{14.55}$$

Hence,

$$P(C|m_1) = p^2$$

Using similar arguments, we have

$$\begin{aligned} P(C|m_2) = P(C|m_4) &= \left[1 - Q\left(\frac{d}{\sqrt{2\mathcal{N}}}\right)\right]\left[1 - 2Q\left(\frac{d}{\sqrt{2\mathcal{N}}}\right)\right] \\ &= p(2p - 1) \end{aligned}$$

and

$$P(C|m_3) = (2p - 1)^2$$

Because of the symmetry of the signals in all four quadrants, we get similar probabilities for the four signals in each quadrant. Hence,

$$\begin{aligned}
P(C) &= \sum_{i=1}^{16} P(C|m_i)P(m_i) \\
&= \frac{1}{16}\sum_{i=1}^{16} P(C|m_i) \\
&= \frac{1}{16}[4p^2 + 4p(2p-1) + 4p(2p-1) + 4(2p-1)^2] \\
&= \frac{1}{4}[9p^2 - 6p + 1] \\
&= \left(\frac{3p-1}{2}\right)^2 \qquad (14.56a)
\end{aligned}$$

and

$$P_{eM} = 1 - P(C) = \frac{9}{4}\left(p + \frac{1}{3}\right)(1-p)$$

In practice, $P_{eM} \to 0$, and, hence, $P(C) \to 1$. This means $p \simeq 1$ [see Eq. (14.56a)], and

$$P_{eM} \simeq 3(1-p) = 3Q\left(\frac{d}{\sqrt{2\mathcal{N}}}\right) \qquad (14.56b)$$

To express this in terms of the received power S_i, we determine $\bar{E}$, the average energy of the signal set in Fig. 14.12. Because E_k, the energy of s_k, is the square of the distance of s_k from the origin,

$$E_1 = \left(\frac{3d}{2}\right)^2 + \left(\frac{3d}{2}\right)^2 = \frac{9}{2}d^2$$

$$E_2 = \left(\frac{3d}{2}\right)^2 + \left(\frac{d}{2}\right)^2 = \frac{5}{2}d^2$$

Similarly,

$$E_3 = \frac{d^2}{2} \qquad \text{and} \qquad E_4 = \frac{5}{2}d^2$$

Hence,

$$\bar{E} = \frac{1}{4}\left[\frac{9}{2}d^2 + \frac{5}{2}d^2 + \frac{d^2}{2} + \frac{5}{2}d^2\right] = \frac{5}{2}d^2$$

and $d^2 = 0.4\bar{E}$. Moreover, for $M = 16$, each symbol carries the information of $\log_2 16 = 4$ bits. Hence, the energy per bit E_b is

$$E_b = \frac{\bar{E}}{4}$$

and

$$\frac{E_b}{\mathcal{N}} = \frac{\bar{E}}{4\mathcal{N}} = \frac{5d^2}{8\mathcal{N}}$$

Hence,

$$P_{eM} = 3Q\left(\frac{d}{\sqrt{2\mathcal{N}}}\right)$$
$$= 3Q\left(\sqrt{\frac{4}{5}\frac{E_b}{\mathcal{N}}}\right) \tag{14.57}$$

A comparison of this with binary PSK [Eq. (13.21b)] shows that 16-point QAM requires almost 2.5 times as much power as does binary PSK; but the rate of transmission is increased by a factor of $\log_2 M = 4$. This comparison does not take into account the fact that P_b, the BER, is somewhat smaller than P_{eM}. In this case, $N = 2$ and $M = 16$. Hence, the receiver in Fig. 14.8 is preferable. Such a receiver is shown in Fig. 14.12c. Note that because all signals are equiprobable,

$$a_i = \frac{-E_i}{2}$$

For QAM, $\varphi_1(t) = \sqrt{2/T_M}\cos\omega_o(t)$ and $\varphi_2(t) = \sqrt{2/T_M}\sin\omega_o(t)$, where $\omega_o = 2\pi/T_M$.

EXAMPLE 14.4 MPSK

Determine the error probability of the optimum receiver for equiprobable MPSK signals, each with energy E.

Figure 14.13a shows the MPSK signal configuration for $M = 8$. Because all the signals are equiprobable, the decision regions are conical, as shown. The message m_1 is transmitted by a signal $s_1(t)$ represented by the vector s_1 $(s_1, 0)$. If the projection in the signal space of the received signal **r** is **q** (q_1, q_2), and the noise is **n** (n_1, n_2), then

$$\mathbf{q} = (s_1 + \mathrm{n}_1,\ \mathrm{n}_2) = (\underbrace{\sqrt{E} + \mathrm{n}_1}_{q_1},\ \underbrace{\mathrm{n}_2}_{q_2})$$

Also,

$$P(C|m_1) = P(\mathbf{q} \text{ lies in } R_1)$$

This is simply the volume under the conical region of the joint PDF of q_1 and q_2. Because n_1 and n_2 are independent gaussian RVs with variance $\mathcal{N}/2$, q_1 and q_2 are independent gaussian variables with means $\sqrt{E}$ and 0, respectively, and each with variance $\mathcal{N}/2$. Hence,

$$p_{\mathrm{q}_1\mathrm{q}_2}(q_1, q_2) = \left[\frac{1}{\sqrt{\pi\mathcal{N}}}e^{-(q_1-\sqrt{E})^2/\mathcal{N}}\right]\left[\frac{1}{\sqrt{\pi\mathcal{N}}}e^{-q_2^2/\mathcal{N}}\right]$$

and

$$P(C|m_1) = \frac{1}{\pi\mathcal{N}}\iint_{R_1} e^{-[(q_1-\sqrt{E})^2+q_2^2]/\mathcal{N}}\,dq_1\,dq_2 \tag{14.58a}$$

$$= \frac{1}{\pi\mathcal{N}}\int_{q_1}\left(\int_{q_2} e^{-q_2^2/\mathcal{N}}\,dq_2\right)e^{-(q_1-\sqrt{E})^2/\mathcal{N}}\,dq_1 \tag{14.58b}$$

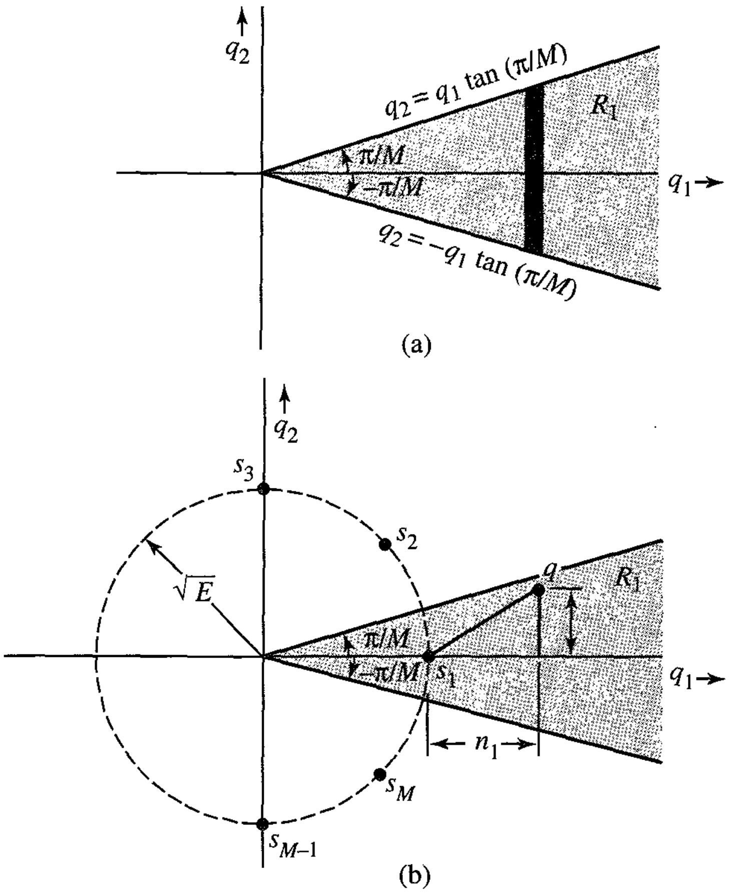


Figure 14.13 MPSK signals.

To integrate over R_1, we first integrate over the (dark) shaded strip in Fig. 14.13b. Along the border of R_1,

$$q_2 = \pm\left(\tan\frac{\pi}{M}\right) q_1$$

Hence,

$$P(C|m_1) = \frac{1}{\pi\mathcal{N}}\int_0^\infty \left(\int_{-q_1\tan(\pi/M)}^{q_1\tan(\pi/M)} e^{-q_2^2/\mathcal{N}}\, dq_2\right) e^{-\left(q_1-\sqrt{E}\right)^2/\mathcal{N}}\, dq_1$$

$$= \frac{1}{\sqrt{\pi\mathcal{N}}}\int_0^\infty \left[1 - 2Q\left(\frac{q_1\tan(\pi/M)}{\sqrt{\mathcal{N}/2}}\right)\right] e^{-\left(q_1-\sqrt{E}\right)^2/\mathcal{N}}\, dq_1$$

Changing the variable to $x = \sqrt{2/\mathcal{N}}\,q_1$, we get

$$P(C|m_1) = \frac{1}{\sqrt{2\pi}}\int_0^\infty \left[1 - 2Q\left(x\tan\frac{\pi}{M}\right)\right] e^{-\left(x-\sqrt{2E/\mathcal{N}}\right)^2/2}\, dx \qquad (14.59a)$$

Using the fact that E_b, the energy per bit, is $E/\log_2 M$, we have

$$P(C|m_1) = \frac{1}{\sqrt{2\pi}}\int_0^\infty \left[1 - 2Q\left(x\tan\frac{\pi}{M}\right)\right] e^{-\left[x-\sqrt{(2\log_2 M)\,E_b/\mathcal{N}}\right]^2/2}\, dx \qquad (14.59b)$$

The integration can also be performed in cylindrical coordinates using the transformation $q_1 = \rho\sqrt{\mathcal{N}/2}\,\cos\theta$ and $q_2 = \rho\sqrt{\mathcal{N}/2}\,\sin\theta$. The limits on ρ are $(0,\ \infty)$ and those on θ are $-\pi/M$ to π/M. Hence,

$$P(C|m_1) = \frac{1}{2\pi}\int_{-\pi/M}^{\pi/M} d\theta \int_0^{\infty} \rho e^{-(\rho^2 - 2\rho\sqrt{2E/\mathcal{N}}\cos\theta + 2E/\mathcal{N})/2}\,d\rho \tag{14.60a}$$

$$= \frac{1}{2\pi}\int_{-\pi/M}^{\pi/M} d\theta \int_0^{\infty} \rho e^{-[\rho^2 - 2\rho\sqrt{(2\log_2 M)\,E_b/\mathcal{N}}\cos\theta + (2\log_2 M)E_b/\mathcal{N}]/2}\,d\rho \tag{14.60b}$$

Because of the symmetry of the signal configuration, $P(C|m_i)$ is the same for all i. Hence,

$$P(C) = P(C|m_1)$$

and

$$P_{eM} = 1 - P(C|m_1)$$

P_{eM} is computed numerically. The plot is shown in Fig. 13.22. Still another form of the integral for P_{eM} was found in Chapter 13 [Eq. (13.53)]. For MPSK, $s_i(t) = \sqrt{2E/T_M}\,\cos\,(\omega_o t + \theta_i)$, where $\omega_o = 2\pi/T_M$, $\theta_i = 2\pi i/M$, and the optimum receiver turns out to be just a phase detector similar to that shown in Fig. 14.12 (see Prob. 14.3-3).

General Expression for Error Probability

Thus far we have considered rather simple schemes where the decision regions can be found easily. The method of computing error probabilities from the knowledge of decision regions has also been discussed. When the dimensions of the signal space increase, it becomes difficult to visualize the decision regions graphically, and as a result the method loses its power. We now develop an analytical expression for computing error probability for a general M-ary scheme.

From the structure of the optimum receiver in Fig. 14.7, we observe that if m_1 is transmitted, then the correct decision will be made only if

$$\mathrm{b}_1 > \mathrm{b}_2,\ \mathrm{b}_3,\ \ldots,\ \mathrm{b}_M$$

In other words,

$$P(C|m_1) = \text{probability } (\mathrm{b}_1 > \mathrm{b}_2,\ \mathrm{b}_3,\ \ldots,\ \mathrm{b}_M|m_1) \tag{14.61}$$

If m_1 is transmitted, then (Fig. 14.7)

$$\mathrm{b}_k = \int_0^{T_M} [s_1(t) + \mathrm{n}(t)]s_k(t)\,dt + a_k \tag{14.62}$$

Let

$$\rho_{ij} = \int_0^{T_M} s_i(t)s_j(t)\,dt \qquad i,\ j = 1,\ 2,\ \ldots,\ M \tag{14.63}$$

ρ_{ij} are known as **crosscorrelation coefficients**. Thus (if m_1 is transmitted),

$$\mathrm{b}_k = \rho_{1k} + \int_0^{T_M} \mathrm{n}(t)s_k(t)\,dt + a_k \tag{14.64a}$$

$$= \rho_{1k} + a_k + \sum_{j=1}^{N} s_{kj}\mathrm{n}_j \tag{14.64b}$$

where n_j is the component of $\mathrm{n}(t)$ along $\varphi_j(t)$. Note that $\rho_{1k} + a_k$ is a constant, and variables n_j $(j = 1, 2, \ldots, N)$ are independent jointly gaussian variables, each with zero mean and a variance of $\mathcal{N}/2$. Thus, variables b_k are a linear combination of jointly gaussian variables. It follows that the variables $\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_M$ are also jointly gaussian. The probability of making a correct decision when m_1 is transmitted can be computed from Eq. (14.61). Note that b_1 can lie anywhere in the range $(-\infty, \infty)$. More precisely, if $p(b_1, b_2, \ldots, b_M|m_1)$ is the joint PDF of $\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_M$, then Eq. (14.61) can be expressed as

$$P(C|m_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{b_1}\cdots\int_{-\infty}^{b_1} p(b_1, b_2, \ldots, b_M|m_1)\, db_1, db_2, \ldots, db_M \tag{14.65a}$$

where the limits of integration of b_1 are $(-\infty, \infty)$, and for the remaining variables the limits are $(-\infty, b_1)$. Thus,

$$P(C|m_1) = \int_{-\infty}^{\infty} db_1 \int_{-\infty}^{b_1} db_2 \ldots \int_{-\infty}^{b_1} p(b_1, b_2, \ldots, b_M|m_1)\, db_M \tag{14.65b}$$

Similarly, $P(C|m_2), \ldots, P(C|m_M)$ can be computed, and

$$P(C) = \sum_{j=1}^{M} P(C|m_j)P(m_j)$$

and

$$P_{eM} = 1 - P(C)$$

EXAMPLE 14.5 Orthogonal Signal Set

In this set all M equal-energy signals $s_1(t), s_2(t), \ldots, s_M(t)$ are mutually orthogonal. As an example, a signal set for $M = 3$ is shown in Fig. 14.14.

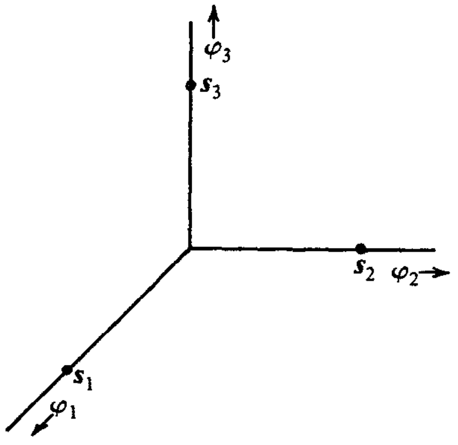


Figure 14.14 Orthogonal signals.

The orthogonal set $\{s_k(t)\}$ is characterized by the fact that

$$\mathbf{s}_j \cdot \mathbf{s}_k = \begin{cases} 0 & j \neq k \\ E & j = k \end{cases} \tag{14.66}$$

Hence,

$$\rho_{ij} = \mathbf{s}_i \cdot \mathbf{s}_j = \begin{cases} 0 & i \neq j \\ E & i = j \end{cases} \tag{14.67}$$

Further, we shall assume all signals to be equiprobable. This yields

$$\begin{aligned} a_k &= \frac{1}{2}\left[\mathcal{N}\ln\left(\frac{1}{M}\right) - E_k\right] \\ &= -\frac{1}{2}(\mathcal{N}\ln M + E) \end{aligned}$$

where E is the energy of each signal. Note that a_k is the same for all values of k. Because the constants a_k enter the expression only for the sake of comparison (see Fig. 14.8b), they can be ignored, and hence, we can let $a_k = 0$. Also for an orthogonal set,

$$s_k(t) = \sqrt{E}\,\varphi_k(t) \tag{14.68}$$

Therefore,

$$s_{kj} = \begin{cases} \sqrt{E} & k = j \\ 0 & k \neq j \end{cases} \tag{14.69}$$

Hence, from Eqs. (14.64b), (14.67), and (14.69), we have (when m_1 is transmitted)

$$\mathrm{b}_k = \begin{cases} E + \sqrt{E}\,\mathrm{n}_1 & k = 1 \\ \sqrt{E}\,\mathrm{n}_k & k = 2, 3, \ldots, M \end{cases} \tag{14.70}$$

Note that $\mathrm{n}_1, \mathrm{n}_2, \ldots, \mathrm{n}_M$ are independent gaussian variables, each with zero mean and variance $\mathcal{N}/2$. Variables b_k that are of the form $(\alpha\mathrm{n}_k + \beta)$ are also independent gaussian variables. Equation (14.70) shows that the variable b_1 has the mean E and variance $(\sqrt{E})^2(\mathcal{N}/2) = \mathcal{N}E/2$. Hence,

$$p_{\mathrm{b}_1}(b_1) = \frac{1}{\sqrt{\pi\mathcal{N}E}} e^{-(b_1-E)^2/\mathcal{N}E}$$

$$p_{\mathrm{b}_k}(b_k) = \frac{1}{\sqrt{\pi\mathcal{N}E}} e^{-b_k^2/\mathcal{N}E} \qquad k = 2, 3, \ldots, M$$

Because $\mathrm{b}_1, \mathrm{b}_2, \ldots, b_M$ are independent, the joint probability density is the product of the individual densities:

$$p(b_1, b_2, \ldots, b_M|m_1) = \frac{1}{\sqrt{\pi\mathcal{N}E}} e^{-(b_1-E)^2/\mathcal{N}E} \prod_{k=2}^{M} \left(\frac{1}{\sqrt{\pi\mathcal{N}E}} e^{-b_k^2/\mathcal{N}E}\right)$$

and

$$
\begin{aligned}
P(C|m_1) &= \frac{1}{\sqrt{\pi \mathcal{N} E}} \int_{-\infty}^{\infty} db_1 [e^{-(b_1-E)^2/\mathcal{N}E}] \times \prod_{k=2}^{M} \left(\int_{-\infty}^{b_1} \frac{1}{\sqrt{\pi \mathcal{N} E}} e^{-b_k^2/\mathcal{N}E}\, db_k \right) \\
&= \frac{1}{\sqrt{\pi \mathcal{N} E}} \int_{-\infty}^{\infty} db_1 [e^{-(b_1-E)^2/\mathcal{N}E}] \times \left(\int_{-\infty}^{b_1} \frac{1}{\sqrt{\pi \mathcal{N} E}} e^{-x^2/\mathcal{N}E}\, dx \right)^{M-1} \\
&= \frac{1}{\sqrt{\pi \mathcal{N} E}} \int_{-\infty}^{\infty} \left[1 - Q\left(\frac{b_1}{\sqrt{\mathcal{N}E/2}} \right) \right]^{M-1} \times e^{-(b_1-E)^2/\mathcal{N}E}\, db_1 \qquad \text{(14.71a)}
\end{aligned}
$$

Changing the variable so that $b_1/\sqrt{\mathcal{N}E/2} = y$, and recognizing that $E/\mathcal{N} = (\log_2 M)$ $E_b/\mathcal{N}$, we obtain

$$
\begin{aligned}
P(C|m_1) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(y-\sqrt{2E/\mathcal{N}}\right)^2/2}\, [1 - Q(y)]^{M-1}\, dy \qquad \text{(14.71b)} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left[y-\sqrt{(2\log_2 M)E_b/\mathcal{N}}\right]^2/2}\, [1 - Q(y)]^{M-1}\, dy \qquad \text{(14.71c)}
\end{aligned}
$$

Note that because this signal set is geometrically symmetrical,

$$P(C|m_1) = P(C|m_2) = \cdots = P(C|m_M)$$

Hence,

$$P(C) = P(C|m_1)$$

and

$$
\begin{aligned}
P_{eM} &= 1 - P(C) \\
&= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left[y-\sqrt{(2\log_2 M)E_b/\mathcal{N}}\right]^2/2}\, [1 - Q(y)]^{M-1}\, dy \qquad \text{(14.71d)}
\end{aligned}
$$

This confirms our earlier result in Eq. (13.57). Figure 13.25 shows the plot of P_{eM} vs. $E_b/\mathcal{N}$.

Significance of the Geometrical Configuration of Signals

From this discussion, one very interesting fact emerges. Whenever the optimum receiver is used, the error probability does not depend on specific signal waveforms but it depends only on their geometrical configuration in the signal space. A given configuration in a signal space can represent infinite possible signal sets, depending on the choice of the basis signals used. A given configuration *does* specify the average energy of the set. This means the error probability depends on signal waveforms only through the average energy of the set. Thus, the average signal energy (or power) emerges as a fundamental parameter that determines the error probability.

This discussion also vividly demonstrates the insight provided by the use of signal space in the study of optimum receivers.

Bandwidth of M-ary Signals

As discussed in Sec. 14.1, the dimensionality of a signal is $2BT_M + 1$, where T_M is the signal duration and B is its essential bandwidth. It follows that for an N-dimensional signal space ($N \leq M$), the bandwidth is $B = (N-1)/2T_M$. Thus, reducing N reduces the bandwidth.

We can verify that N-dimensional signals can be transmitted over $(N-1)/2T_M$ Hz by constructing a specific signal set. Let us choose the signals

$$\begin{aligned}
\varphi_0(t) &= \frac{1}{\sqrt{T_M}} \\
\varphi_1(t) &= \sqrt{\frac{2}{T_M}}\sin \omega_o t \\
\varphi_2(t) &= \sqrt{\frac{2}{T_M}}\cos \omega_o t \qquad \omega_o = \frac{2\pi}{T_M} \\
\varphi_3(t) &= \sqrt{\frac{2}{T_M}}\sin 2\omega_o t \qquad 0 \leq t \leq T_M \qquad\qquad (14.72) \\
\varphi_4(t) &= \sqrt{\frac{2}{T_M}}\cos 2\omega_o t \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\varphi_{k-1}(t) &= \sqrt{\frac{2}{T_M}}\sin\left(\frac{k}{2}\omega_o t\right) \\
\varphi_k(t) &= \sqrt{\frac{2}{T_M}}\cos\left(\frac{k}{2}\omega_o t\right)
\end{aligned}$$

These are $k+1$ orthogonal pulses with a total bandwidth of $(k/2)(\omega_o/2\pi) = k/2T_M$ Hz. Hence, when $k+1 = N$, the bandwidth* is $(N-1)/2T_M$. Thus, $N = 2T_M B + 1$.

To attain a given error probability, there is a trade-off between the average energy of the signal set and its bandwidth. If we reduce the signal space dimensionality, the transmission bandwidth is reduced. But the signals are now closer together, because of the reduced dimensionality. This will increase P_{eM}. Hence, to maintain a given P_{eM}, we must move the signals farther apart; that is, increase energy. Thus, the cost of reduced bandwidth is paid in terms of increased energy. In Chapter 15 we shall study the ideal trade-off between bandwidth and signal energy and compare the performance of various signal schemes in light of this relationship.

---

* Here we are ignoring the band spreading at the edge. This spread is about $1/T_M$ Hz. The actual bandwidth is larger than $(N-1)/2T_M$ by this amount.

14.4 EQUIVALENT SIGNAL SETS

The computation of error probabilities is greatly facilitated by the translation and rotation of coordinate axes. We now show that such operations are permissible.

Consider a signal set with its corresponding decision regions, as shown in Fig. 14.15a. The conditional probability $P(C|m_1)$ is the probability that the noise vector drawn from s_1 lies within R_1. Note that this probability does not depend on the origin of the coordinate system. We may translate the coordinate system any way we wish. This is equivalent to translating the signal set and the corresponding decision regions. Thus, the $P(C|m_i)$ for the translated system shown in Fig. 14.15b is identical to that of the system in Fig. 14.15a.

In the case of gaussian noise, we make another important observation. The rotation of the coordinate system does not affect the error probability because the noise-vector probability

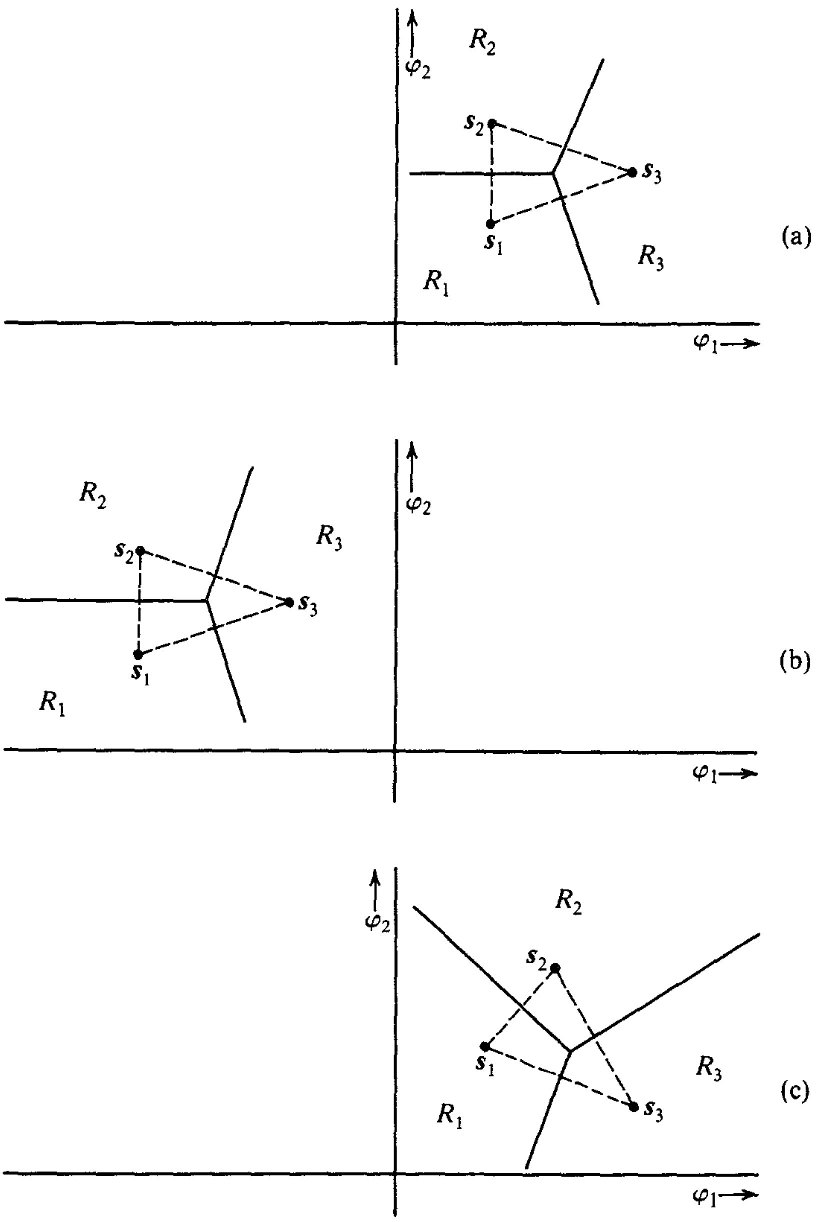


Figure 14.15 Translation and rotation of coordinate axes.

density has spherical symmetry. To show this we shall consider Fig. 14.15c, where the signal set in Fig. 14.15a is shown translated and rotated. Note that a rotation of the coordinate system is equivalent to a rotation of the signal set in the opposite sense. Here for convenience we rotate the signal set instead of the coordinate system. It can be seen that the probability that the noise vector **n** drawn from s_1 lies in R_1 is the same in Fig. 14.15a and c, because this probability is given by the integral of the noise probability density $p_{\mathbf{n}}(\boldsymbol{n})$ over the region R_1. Because $p_{\mathbf{n}}(\boldsymbol{n})$ has a spherical symmetry for gaussian noise, the probability will remain unaffected by a rotation of the region R_1. Clearly, for additive gaussian channel noise, translation and rotation of the coordinate system (or translation and rotation of the signal set) do not affect the error probability. Note that when we rotate or translate a set of signals, the resulting set represents an entirely different set of signals. Yet the error probabilities of the two sets are identical. Such sets are called **equivalent sets.**

The following example demonstrates the utility of translation and rotation of a signal set in the computation of error probability.

EXAMPLE 14.6 **A quaternary PSK (QPSK)** signal set is shown in Fig. 14.16a,

$$s_1 = -s_2 = \sqrt{E}\,\varphi_1$$

$$s_3 = -s_4 = \sqrt{E}\,\varphi_2$$

Assuming all symbols to be equiprobable, determine P_{eM} for an AWGN channel with noise PSD $\mathcal{N}/2$.

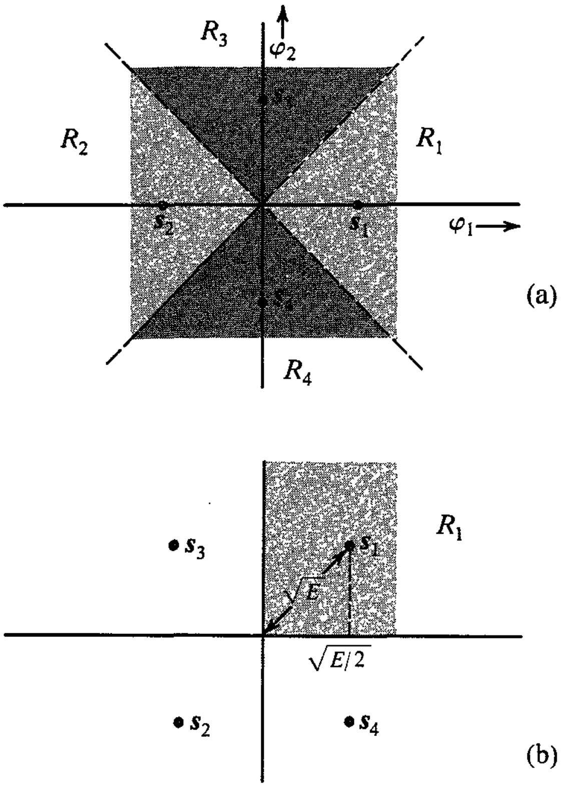


Figure 14.16 Analysis of QPSK.

This problem has already been solved in Example 14.4 for a general value of M. Here we shall solve it for $M = 4$ to demonstrate the power of the rotation of axes.

Because all the symbols are equiprobable, the decision region boundaries will be perpendicular bisectors of lines joining various signal points (Fig. 14.16a). Now

$$P(C|m_1) = P(\text{noise vector originating at } s_1 \text{ remains in } R_1) \tag{14.73}$$

This can be found by integrating the joint PDF of components n_1 and n_2 (originating at s_1) over the region R_1. This double integral can be found by using suitable limits, as in Eq. (14.59). The problem is greatly simplified, however, if we rotate the signal set by 45°, as shown in Fig. 14.16b. The decision regions are rectangular, and if n_1 and n_2 are noise components along φ_1 and φ_2, then Eq. (14.73) can be expressed as

$$\begin{aligned}
P(C|m_1) &= P\left(\mathrm{n}_1 > -\sqrt{\frac{E}{2}},\ \mathrm{n}_2 > -\sqrt{\frac{E}{2}}\right) \\
&= P\left(\mathrm{n}_1 > -\sqrt{\frac{E}{2}}\right) P\left(\mathrm{n}_2 > -\sqrt{\frac{E}{2}}\right) \\
&= \left[1 - Q\left(\sqrt{\frac{E}{2\sigma_n^2}}\right)\right]^2 \\
&= \left[1 - Q\left(\sqrt{\frac{E}{\mathcal{N}}}\right)\right]^2 && \text{(14.74a)} \\
&= \left[1 - Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right)\right]^2 && \text{(14.74b)}
\end{aligned}$$

Minimum-Energy Signal Set

As noted earlier, an infinite number of possible equivalent signal sets exist. Because signal energy depends on its distance from the origin, however, equivalent sets do not necessarily have the same average energy. Thus, among the infinite possible equivalent signal sets, the one in which the signals are closest to the origin has the minimum average signal energy (or transmitted power).

Let $m_1, m_2, \ldots, m_M$ be M messages with waveforms $s_1(t), s_2(t), \ldots, s_M(t)$ represented by points $s_1, s_2, \ldots, s_M$ in the signal space. The mean energy of these signals is $\bar{E}$, given by

$$\bar{E} = \sum_{i=1}^{M} P(m_1)|s_i|^2$$

Translation of this signal set is equivalent to subtracting some vector $\boldsymbol{a}$ from each signal. Let this operation yield the minimum-mean-energy set. We now wish to find the vector $\boldsymbol{a}$ such that the new mean energy

$$\bar{E} = \sum_{i=1}^{M} P(m_i)|\boldsymbol{s}_i - \boldsymbol{a}|^2 \tag{14.75}$$

is minimum. We can show that $\boldsymbol{a}$ must be the center of gravity of M points located at $s_1, s_2, \ldots, s_M$ with masses $P(m_1), P(m_2), \ldots, P(m_M)$, respectively:

$$\boldsymbol{a} = \sum_{i=1}^{M} P(m_i)\boldsymbol{s}_i = \overline{\boldsymbol{s}_i} \tag{14.76}$$

To prove this, suppose the mean energy is minimum for some translation $\boldsymbol{b}$. Then

$$\begin{aligned}\bar{E} &= \sum_{i=1}^{M} P(m_i)|\boldsymbol{s}_i - \boldsymbol{b}|^2 \\ &= \sum_{i=1}^{M} P(m_i)|(\boldsymbol{s}_i - \boldsymbol{a}) + (\boldsymbol{a} - \boldsymbol{b})|^2 \\ &= \sum_{i=1}^{M} P(m_i)|\boldsymbol{s}_i - \boldsymbol{a}|^2 + 2(\boldsymbol{a} - \boldsymbol{b}) \cdot \sum_{i=1}^{M} P(m_i)(\boldsymbol{s}_i - \boldsymbol{a}) + \sum_{i=1}^{M} P(m_i)|\boldsymbol{a} - \boldsymbol{b}|^2\end{aligned}$$

Observe that the second term in the above expression vanishes because of the relationship in Eq. (14.76). Hence,

$$\bar{E} = \sum_{i=1}^{M} P(m_i)|\boldsymbol{s}_i - \boldsymbol{a}|^2 + \sum_{i=1}^{M} P(m_i)|\boldsymbol{a} - \boldsymbol{b}|^2$$

This is minimum when $\boldsymbol{b} = \boldsymbol{a}$. Note that the rotation of the coordinates does not change the energy, and, hence, there is no need to rotate the signal set to minimize the energy.

EXAMPLE 14.7 For the binary orthogonal signal set shown in Fig. 14.17a, determine the minimum-energy equivalent signal set.

The minimum-energy set for this case is shown in Fig. 14.17b. The origin lies at the center of gravity of the signals. We have also rotated the signals for convenience. The distances k_1 and k_2 must be such that

$$k_1 + k_2 = d$$

and

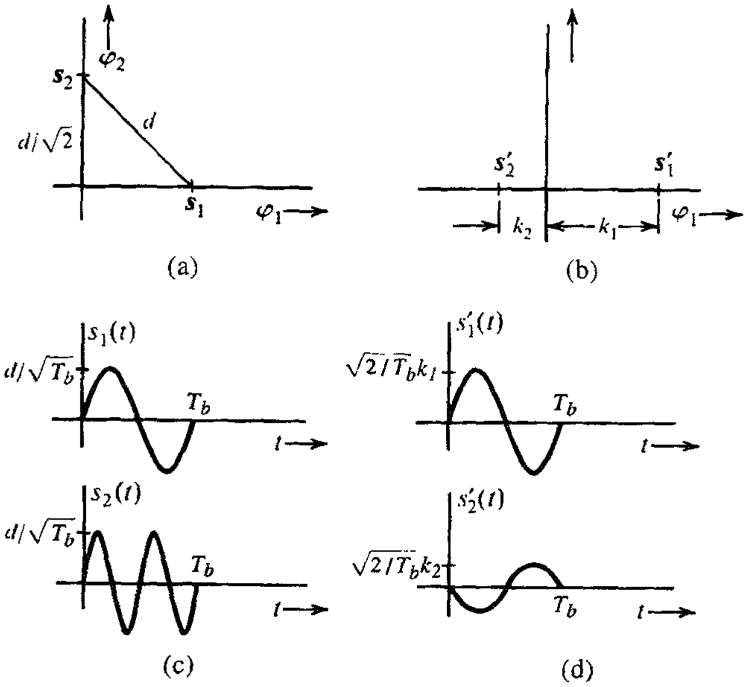
$$k_1 P(m_1) = k_2 P(m_2)$$

Solution of these two equations yields

$$k_1 = P(m_2)d$$

and

$$k_2 = P(m_1)d$$


φ_2
s_2
$d/\sqrt{2}$
d
s_1
φ_1
s_2'
s_1'
k_2
k_1
φ_1
$s_1(t)$
$d/\sqrt{T_b}$
T_b
t
$s_1'(t)$
$\sqrt{2/T_b}k_1$
T_b
t
$s_2(t)$
$d/\sqrt{T_b}$
T_b
t
$s_2'(t)$
$\sqrt{2/T_b}k_2$
T_b
t

(a) (b) (c) (d)

Figure 14.17 Equivalent signal sets.

Both signal sets (Fig. 14.17a and b) have the same error probability, but the latter has a smaller mean energy. If $\bar{E}$ and $\overline{E'}$ are the respective mean energies of the two sets, then

$$\bar{E} = P(m_1)\frac{d^2}{2} + P(m_2)\frac{d^2}{2} = \frac{d^2}{2}$$

and

$$\begin{aligned}\overline{E'} &= P(m_1)k_1^2 + P(m_2)k_2^2 \\ &= P(m_1)P^2(m_2)d^2 + P(m_2)P^2(m_1)d^2 \\ &= P(m_1)P(m_2)d^2\end{aligned}$$

Note that the product $P(m_1)P(m_2)$ is maximum when $P(m_1) = P(m_2) = 1/2$, in which case

$$P(m_1)P(m_2) = \frac{1}{4}$$

and

$$\overline{E'} \le \frac{d^2}{4}$$

Therefore,

$$\overline{E'} \le \frac{\bar{E}}{2}$$

and for the case of equiprobable signals,

$$\overline{E'} = \frac{\bar{E}}{2}$$

In this case,

$$k_1 = k_2 = \frac{d}{2}$$

$$\bar{E} = \frac{d^2}{2} \quad \text{and} \quad \overline{E'} = \frac{d^2}{4}$$

The signals in Fig. 14.17b are called **antipodal signals** when $k_1 = k_2$. The error probability of the signal set in Fig. 14.17a (and Fig. 14.17b) is equal to that in Fig. 14.11a and can be found from Eq. (14.54a).

As a concrete example, let us choose the basis signals as sinusoids of frequency $\omega_o = 2\pi/T_M$:

$$\left.\begin{aligned} \varphi_1(t) &= \sqrt{\frac{2}{T_M}} \sin \omega_o t \\ \varphi_2(t) &= \sqrt{\frac{2}{T_M}} \sin 2\omega_o t \end{aligned}\right\} \quad 0 \le t < T_M$$

Hence,

$$\left.\begin{aligned} s_1(t) &= \frac{d}{\sqrt{2}}\varphi_1(t) = \frac{d}{\sqrt{T_M}} \sin \omega_o t \\ s_2(t) &= \frac{d}{\sqrt{2}}\varphi_2(t) = \frac{d}{\sqrt{T_M}} \sin 2\omega_o t \end{aligned}\right\} \quad 0 \le t < T_M$$

The signals $s_1(t)$ and $s_2(t)$ are shown in Fig. 14.17c, and the geometrical representation is shown in Fig. 14.17a. Both signals are located at a distance $d/\sqrt{2}$ from the origin, and the distance between the signals is d.

The minimum energy signals $s_1'(t)$ and $s_2'(t)$ for this set are given by

$$\left.\begin{aligned} s_1'(t) &= \sqrt{\frac{2}{T_M}} P(m_2) d \, \sin \omega_o t \\ s_2'(t) &= \sqrt{\frac{2}{T_M}} P(m_1) d \, \sin \omega_o t \end{aligned}\right\} \quad 0 \le t < T_M$$

These signals are sketched in Fig. 14.17d.

EXAMPLE 14.8 Simplex Signal Set

A minimum-energy equivalent set of an equiprobable orthogonal set is called a **simplex**, or **transorthogonal, signal set**. Derive the simplex set corresponding to the orthogonal set in Eq. (14.66).

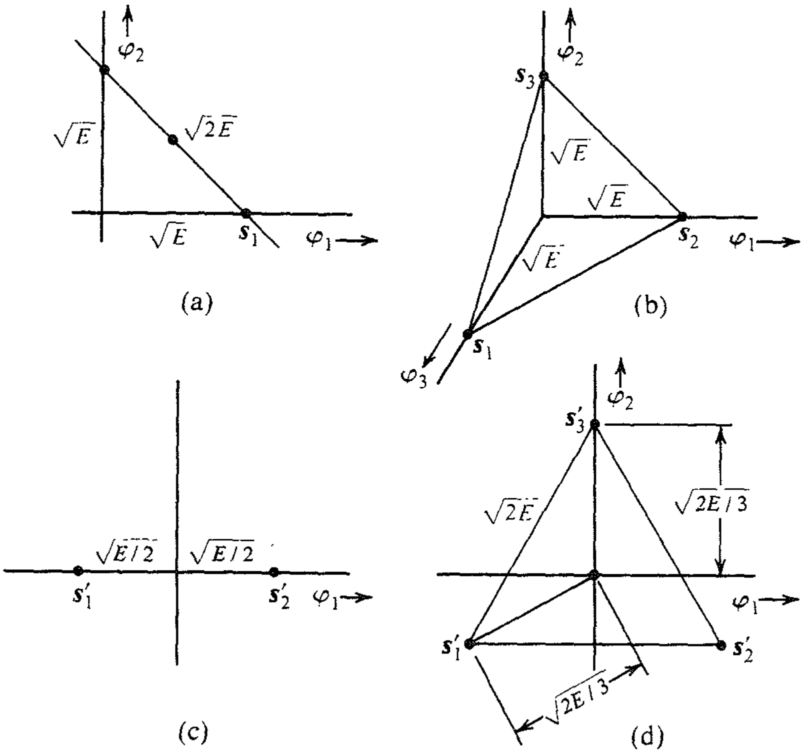

φ_2
$\sqrt{E}$
$\sqrt{2E}$
$\sqrt{E}$
s_1
φ_1
(a)
φ_2
s_3
$\sqrt{E}$
$\sqrt{E}$
$\sqrt{E}$
s_2
φ_1
s_1
φ_3
(b)
$\sqrt{E/2}$
$\sqrt{E/2}$
s'_1
s'_2
φ_1
(c)
φ_2
s'_3
$\sqrt{2E}$
$\sqrt{2E/3}$
φ_1
s'_1
s'_2
$\sqrt{2E/3}$
(d)

Figure 14.18 Simplex signals.

To obtain the minimum-energy set, the origin should be shifted to the center of gravity of the signal set. For the 2-dimensional case (Fig. 14.18a), the simplex set is shown in Fig. 14.18c, and for the 3-dimensional case (Fig. 14.18b), the simplex set is shown in Fig. 14.18d. Note that the dimensionality of the simplex signal set is less than that of the orthogonal set by 1. This is true in general for any value of M. It can be shown that the simplex signal set is the optimum (minimum error probability) for the case of equiprobable signals embedded in white gaussian noise when energy is constrained.[6,7]

We can calculate the mean energy of the simplex set by noting that it is obtained by translating the orthogonal set by a vector $\boldsymbol{a}$, given by Eq. (14.76),

$$\boldsymbol{a} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{s}_i$$

For orthogonal signals,

$$\boldsymbol{s}_i = \sqrt{E}\,\boldsymbol{\varphi}_i$$

Therefore,

$$a = \frac{\sqrt{E}}{M} \sum_{i=1}^{M} \varphi_i$$

where E is the energy of each signal in the orthogonal set and φ_i is the unit vector along the ith coordinate axis. The signals in the simplex set are given by

$$\begin{aligned} s_k' &= s_k - a \\ &= \sqrt{E}\varphi_k - \frac{\sqrt{E}}{M} \sum_{i-1}^{M} \varphi_i \end{aligned} \tag{14.77}$$

The energy E' of signal s_k' is given by $|s_k'|^2$,

$$E' = s_k' \cdot s_k' \tag{14.78}$$

Substituting Eq. (14.77) into Eq. (14.78) and observing that the set φ_i is orthonormal, we have

$$\begin{aligned} E' &= E - \frac{E}{M} \\ &= E\left(1 - \frac{1}{M}\right) \end{aligned} \tag{14.79}$$

Hence, for the same performance (error probability), the mean energy of the simplex signal set is $1 - 1/M$ times that of the orthogonal signal set. For $M \gg 1$, the difference is not significant. For this reason and because they are easier to generate, orthogonal rather than simplex signals are used in practice whenever M exceeds 4 or 5.

In the next chapter, we shall show that in the limit as $M \to \infty$, the orthogonal (as well as the simplex) signals attain the upper bound of performance predicted by Shannon's theorem.

14.5 NONWHITE (COLORED) CHANNEL NOISE

Thus far we have restricted our discussion exclusively to white gaussian channel noise. Our discussion can be extended with relative ease to nonwhite, or colored, gaussian channel noise.

If the noise PSD $S_n(\omega)$ is not white, we use a noise-whitening filter $H(\omega)$ at the input of the receiver, where

$$|H(\omega)| = \frac{1}{\sqrt{S_n(\omega)}}$$

Consider a signal set $\{s_i(t)\}$ and a channel noise $n(t)$ that is not white [$S_n(\omega)$ is not constant]. At the input of the receiver, we use a noise-whitening filter $H(\omega)$ that transforms the colored noise into white noise (Fig. 14.19). But it also alters the signal set $\{s_i(t)\}$ to $\{s_i'(t)\}$, where

$$s_i'(t) = s_i(t) * h(t)$$

or

$$S_i'(\omega) = S_i(\omega)H(\omega)$$

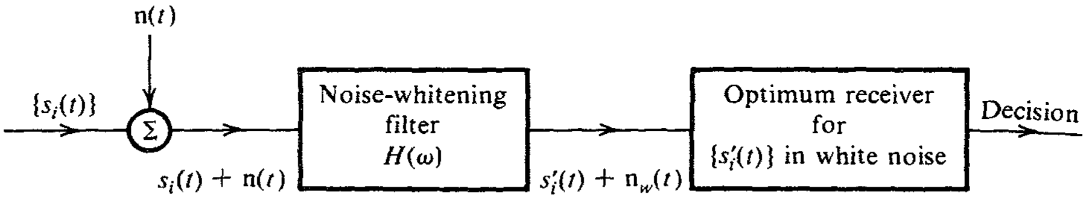

n(t)
{s_i(t)}
Σ
s_i(t) + n(t)
Noise-whitening filter H(ω)
s'_i(t) + n_w(t)
Optimum receiver for {s'_i(t)} in white noise
Decision

Figure 14.19 Optimum M-ary receiver for nonwhite channel noise.

We now have a new signal set $\{s_i'(t)\}$ mixed with white gaussian noise, for which the optimum receiver and the corresponding error probability can be determined by the method discussed earlier.

In general, a whitening filter may be unrealizable. In such a case we need to allow a sufficient delay in $H(\omega)$ to be able to closely realize the filter.

14.6 OTHER USEFUL PERFORMANCE CRITERIA

The optimum receiver uses the decision strategy that makes the best possible use of the observed data and any a priori information available. The strategy will also depend on the weights assigned to various types of errors. In this chapter we have thus far assumed that all errors have equal weight (or equal cost). This assumption may not be justified in some cases, and we may therefore have to alter the decision rule.

Generalized Bayes Receiver

If we are given a priori probabilities and the cost functions of various types of errors, the receiver that minimizes the average cost of decision is called the **Bayes receiver**, and the decision rule is **Bayes' decision rule**. Note that the receiver that has been discussed so far is the Bayes receiver under the condition that all errors have equal cost (equal weight). To generalize this rule, let

$$C_{kj} = \text{cost of deciding that } \hat{m} = m_k \text{ when } m_j \text{ was transmitted} \tag{14.80}$$

and, as usual,

$$P(m_i|\mathbf{q}) = \text{conditional probability that } m_i \text{ was transmitted when } \mathbf{q} \text{ is received}$$

If $\mathbf{q}$ is received, then the probability that m_j was transmitted is $P(m_j|\mathbf{q})$ for all $j = 1, 2, \ldots, M$. Hence, the average cost of the decision $\hat{m} = m_k$ is β_k, given by

$$\begin{aligned}\beta_k &= C_{k1}P(m_1|\mathbf{q}) + C_{k2}P(m_2|\mathbf{q}) + \cdots + C_{kM}(m_M|\mathbf{q})\\ &= \sum_{j=1}^{M} C_{kj}P(m_j|\mathbf{q})\end{aligned} \tag{14.81}$$

Thus, if $\mathbf{q}$ is received, the optimum receiver decides that $\hat{m} = m_k$ if

$$\beta_k < \beta_i \qquad \text{for all } i \neq k$$

or

$$\sum_{j=1}^{M} C_{kj} P(m_j|\boldsymbol{q}) < \sum_{j=1}^{M} C_{ij} P(m_j|\boldsymbol{q}) \qquad \text{for all } i \neq k \tag{14.82}$$

Use of Bayes' mixed rule in Eq. (14.82) yields

$$\sum_{j=1}^{M} C_{kj} P(m_j) p_{\mathbf{q}}(\boldsymbol{q}|m_j) < \sum_{j=1}^{M} C_{ij} P(m_j) p_{\mathbf{q}}(\boldsymbol{q}|m_j) \qquad \text{for all } i \neq k \tag{14.83}$$

Note that C_{kk} is the cost of setting $\hat{m} = m_k$ when m_k is transmitted. This cost is generally zero. If we assign equal weight to all other errors, then

$$C_{kj} = \begin{cases} 0 & k = j \\ 1 & k \neq j \end{cases}$$

and the decision rule in Eq. (14.83) reduces to the rule in Eq. (14.42), as expected. The generalized Bayes receiver for $M = 2$, assuming $C_{11} = C_{22} = 0$, sets $\hat{m} = m_1$ if

$$C_{12} P(m_2) p_{\mathbf{q}}(\boldsymbol{q}|m_2) < C_{21} P(m_1) p_{\mathbf{q}}(\boldsymbol{q}|m_1)$$

Otherwise, the receiver decides that $\hat{m} = m_2$.

Maximum-Likelihood Receiver

The strategy used in the Bayes receiver discussed in the preceding subsection is general, except that it can be implemented only when the a priori probabilities $P(m_1)$, $P(m_2)$, ..., $P(m_M)$ are known. Frequently this information is not available. Under these conditions various possibilities exist, depending on the assumptions made. When, for example, there is no reason to expect any one signal to be more likely than any other, we may assign equal probabilities to all the messages:

$$P(m_1) = P(m_2) = \cdots = P(m_M) = \frac{1}{M}$$

Bayes' rule [Eq. (14.42)] in this case becomes: set $\hat{m} = m_k$ if

$$p_{\mathbf{q}}(\boldsymbol{q}|m_k) > p_{\mathbf{q}}(\boldsymbol{q}|m_i) \qquad \text{for all } i \neq k \tag{14.84}$$

Observe that $p_{\mathbf{q}}(\boldsymbol{q}|m_k)$ represents the probability of observing $\boldsymbol{q}$ when m_k is transmitted. Thus, the receiver chooses that signal which, when transmitted, will maximize the likelihood (probability) of observing the received $\boldsymbol{q}$. Hence, this receiver is called the **maximum-likelihood receiver**. Note that the maximum-likelihood receiver is a Bayes receiver under the condition that the a priori message probabilities are equal. In terms of geometrical concepts, the maximum-likelihood receiver decides in favor of that signal which is closest to the received data $\boldsymbol{q}$. The practical implementation of the maximum-likelihood receiver is the same as that of the Bayes receiver (Figs. 14.7 and 14.8) under the condition that all a priori probabilities are equal to $1/M$.

If the signal set is geometrically symmetrical, and if all a priori probabilities are equal (maximum-likelihood receiver), then the decision regions for various signals are congruent.

In this case, because of symmetry, the conditional probability of a correct decision is the same no matter which signal is transmitted, that is,

$$P(C|m_i) = \text{constant} \qquad \text{for all } i$$

Because

$$P(C) = \sum_{i=1}^{M} P(m_i)P(C|m_i)$$

in this case

$$P(C) = P(C|m_i) \tag{14.85}$$

Thus, the error probability of the maximum-likelihood receiver is independent of the actual source statistics $P(m_i)$ for the case of symmetrical signal sets. It should, however, be realized that if the actual source statistics were known beforehand, one could design a better receiver using Bayes' decision rule.

It is apparent that if the source statistics are not known, the maximum-likelihood receiver proves very attractive for a symmetrical signal set. In such a receiver one can specify the error probability independently of the actual source statistics.

Minimax Receiver

Designing a receiver with a certain decision rule completely specifies the conditional probabilities $P(C|m_i)$. The probability of error is given by

$$\begin{aligned} P_{eM} &= 1 - P(C) \\ &= 1 - \sum_{i=1}^{M} P(m_i)P(C|m_i) \end{aligned}$$

Thus, in general, for a given receiver (with some specified decision rule) the error probability depends on the source statistics $P(m_i)$. The error probability is the largest for some source statistics. The error probability in the worst possible case is $[P_{eM}]_{\max}$ and represents the upper bound on the error probability of the given receiver. This upper bound $[P_{eM}]_{\max}$ serves as an indication of the quality of the receiver. Each receiver (with a certain decision rule) will have a certain $[P_{eM}]_{\max}$. The receiver that has the smallest upper bound on the error probability, that is, the minimum $[P_{eM}]_{\max}$, is called the **minimax receiver**.

We shall illustrate the minimax concept for a binary receiver with on-off signaling. The conditional PDFs of the receiving-filter output sample r at $t = T_b$ are $p(r|\mathbf{1})$ and $p(r|\mathbf{0})$. These are the PDFs of r for the "on" and the "off" pulse (i.e., no pulse), respectively. Figure 14.20a shows these PDFs with a certain threshold a. If we receive r $\geq a$, we choose the hypothesis "signal present" (**1**), and the shaded area to the right of a is the probability of **false alarm** (deciding "signal present" when in fact the signal is not present). If r $< a$, we choose the hypothesis "signal absent" (**0**), and the shaded area to the left of a is the probability of **false dismissal** (deciding "signal absent" when in fact the signal is present). It is obvious that the larger is the threshold a, the larger is the false dismissal error and the smaller is the false alarm error (see Fig. 14.20b).

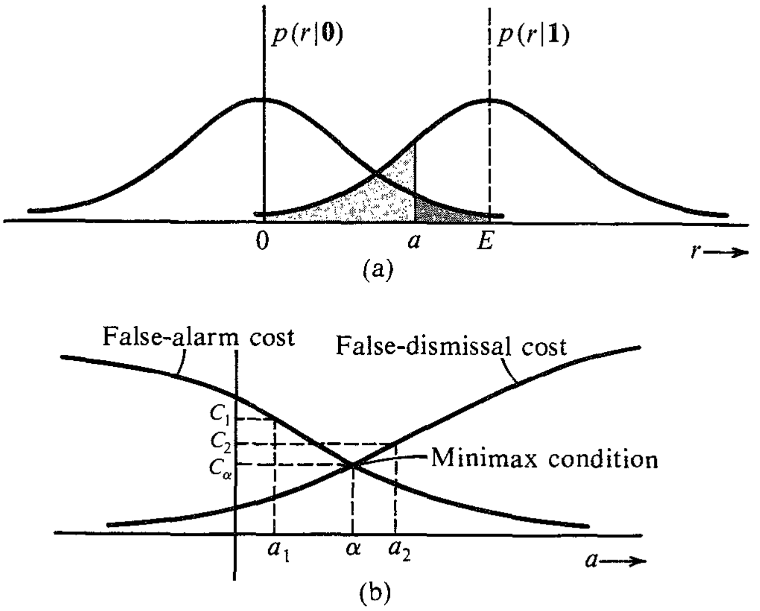


Figure 14.20 Explanation of minimax concept.

We shall now find the minimax condition for this receiver. For the minimax receiver, we consider all possible receivers (all possible values of a in this case) and find the maximum error probability (or cost) that occurs under the worst possible a priori probability distribution. Let us choose $a = a_1$, as shown in Fig. 14.20b. In this case the worst possible case occurs when $P(\mathbf{0}) = 1$ and $P(\mathbf{1}) = 0$, that is, when the signal $s(t)$ is always absent. The type of error in this case is false alarm. These errors have a cost C_1. On the other hand, if we choose $a = a_2$, the worst possible case occurs when $P(\mathbf{0}) = 0$ and $P(\mathbf{1}) = 1$, that is, when the signal is always present, causing only the false-dismissal type of errors. These errors have a cost C_2. It is evident that for the setting $a = \alpha$, the cost of false alarm and false dismissal are equal, namely, C_α. Hence, for all possible source statistics the cost is C_α. Because $C_\alpha < C_1$ and C_2, this cost is the *minimum* of the maximum possible cost (because the worst cases are considered) that accrues for all values of a. Hence, $a = \alpha$ represents the minimax setting.

It follows from this discussion that the minimax receiver is rather conservative. It is designed under the pessimistic assumption that the worst possible source statistics exist. The maximum-likelihood receiver, on the other hand, is designed on the assumption that all messages are equally likely. It can, however, be shown that for a symmetrical signal set, the maximum-likelihood receiver is in fact the minimax receiver. This can be proved by observing that for a symmetrical set, the probability of error of a maximum-likelihood receiver (equal a priori probabilities) is independent of the source statistics [Eq. (14.85)]. Hence, for a symmetrical set, the error probability $P_{eM} = \alpha$ of a maximum-likelihood receiver is also equal to its $[P_{eM}]_{\max}$. We now show that no other receiver exists whose $[P_{eM}]_{\max}$ is less than the α of a maximum-likelihood receiver for a symmetrical signal set. This is seen from the fact that for equiprobable messages, the maximum-likelihood receiver is optimum by definition. All other receivers must have $P_{eM} > \alpha$ for equiprobable messages. Hence, $[P_{eM}]_{\max}$ for these receivers can never be less than α. This proves that the maximum-likelihood receiver is indeed the minimax receiver for a symmetrical signal set.

REFERENCES

1. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty, III: The Dimensions of Space of Essentially Time- and Band-Limited Signals," *Bell Syst. Tech. J.*, vol. 41, pp. 1295–1336, July 1962.
2. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
3. E. Arthurs and H. Dym, "On Optimum Detection of Digital Signals in the Presence of White Gaussian Noise—A Geometric Interpretation and a Study of Three Basic Data Transmission Systems," *IRE Trans. Commun. Syst.*, vol. CS-10, pp. 336–372, Dec. 1962.
4. B. P. Lathi, *An Introduction to Random Signals and Communication Theory*, International Textbook Co., Scranton, PA, 1968.
5. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vols. I, II, and III, Wiley, New York, 1968–1971.
6. H. J. Landau and D. Slepian, "On the Optimality of the Regular Simplex Code," *Bell Syst. Tech. J.*, vol. 45, pp. 1247–1272, Oct. 1966.
7. A. V. Balakrishnan, "Contribution to the Sphere-Packing Problem of Communication Theory," *J. Math. Anal. Appl.*, vol. 3, pp. 485–506, Dec. 1961.

PROBLEMS

14.1-1 Find at least three different sets of orthonormal basis signals for a 5-dimensional signal space.

14.1-2 The basis signals of a 3-dimensional signal space are given by $\varphi_1(t) = p(t)$, $\varphi_2(t) = p(t - T_o)$, and $\varphi_3(t) = p(t - 2T_o)$, where

$$p(t) = \frac{1}{\sqrt{T_o}}[u(t) - u(t - T_o)]$$

Sketch the waveforms of the signals represented by $(1,\ 1,\ 0)$, $(2,\ -1,\ 1)$, $(3,\ 2,\ -\frac{1}{2})$, and $(-\frac{1}{2},\ -1,\ 1)$ in this space.

14.1-3 Repeat Prob. 14.1-2 if

$$\begin{aligned} \varphi_1(t) &= \frac{1}{\sqrt{T_o}} \\ \varphi_2(t) &= \sqrt{\frac{2}{T_o}} \sin \frac{2\pi}{T_o}t \qquad\qquad 0 \le t \le T_o \\ \varphi_3(t) &= \sqrt{\frac{2}{T_o}} \cos \frac{2\pi}{T_o}t \end{aligned}$$

14.1-4 If $p(t)$ is as in Prob. 14.1-2 and

$$\varphi_k(t) = p[t - (k-1)T_o] \qquad k = 1, 2, 3, 4, 5$$

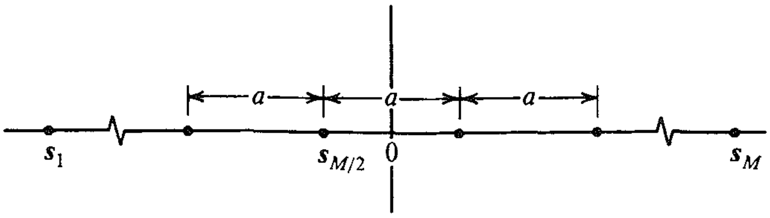
(a) Sketch the signals represented by $(-1,\ 2,\ 3,\ 1,\ 4)$, $(2,\ 1,\ -4,\ -4,\ 2)$, $(3,\ -2,\ 3,\ 4,\ 1)$, and $(-2,\ 4,\ 2,\ 2,\ 0)$ in this space.

(b) Find the energy of each signal.

(c) Find the pairs of signals that are orthogonal.

14.2-1 For a certain stationary gaussian random process $x(t)$, it is given that $R_x(\tau) = e^{-|\tau|}$. Determine the joint PDF of RVs $x(t)$, $x(t+1)$, and $x(t+2)$.

14.3-1 A source emits M equiprobable messages, which are assigned signals $s_1, s_2, \ldots, s_M$, as shown in Fig. P14.3-1. Determine the optimum receiver and the corresponding error probability P_{eM} for an AWGN channel as a function of E_b.



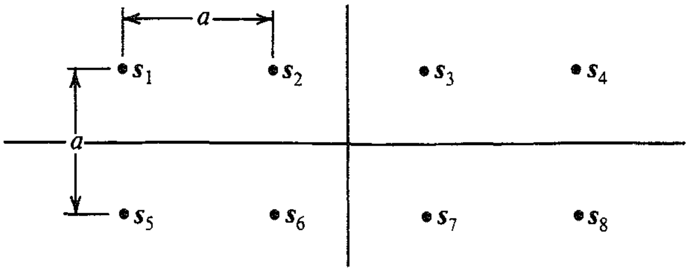


Figure P14.3-1 / Figure P14.3-2

14.3-2 A source emits eight equiprobable messages, which are assigned signals $s_1, s_2, \ldots, s_8$, as shown in Fig. P14.3-2.

(a) Find the optimum receiver for an AWGN channel.

(b) Determine the decision regions and the error probability P_{eM} of the optimum receiver as a function of E_b.

14.3-3 Show that for MPSK, the optimum receiver of the form in Fig. 14.8a is equivalent to a phase comparator. Assume all messages equiprobable and an AWGN channel.

14.3-4 The vertices of an N-dimensional hypercube are a set of 2^N signals

$$s_k(t) = \frac{d}{2}\sum_{j=1}^{N} a_{kj}\varphi_j(t)$$

where $\{\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t)\}$ is a set of N orthonormal signals, and a_{kj} is either 1 or -1. Note that all the N signals are at a distance of $\sqrt{N}d/2$ from the origin and form the vertices of the N-dimensional cube.

(a) Sketch the signal configuration in the signal space for $N = 1, 2$, and 3.

(b) For each configuration in part **(a)**, sketch one possible set of waveforms.

(c) If all the 2^N symbols are equiprobable, find the optimum receiver and determine the error probability P_{eM} of the optimum receiver as a function of E_b assuming an AWGN channel.

14.3-5 A ternary signal configuration is shown in Fig. P14.3-5.

(a) If $P(m_o) = 0.5$ and $P(m_1) = P(m_{-1}) = 0.25$, determine the optimum decision regions and P_{eM} of the optimum receiver as a function of $\bar{E}$. Assume an AWGN channel.

(b) Find P_{eM} as a function of $\bar{E}/\mathcal{N}$.

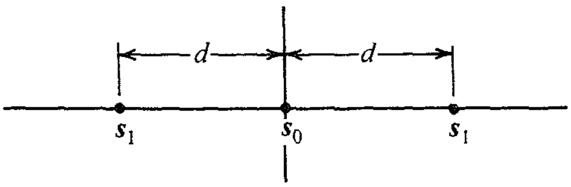


Figure P14.3-5

14.3-6 A 16-ary signal configuration is shown in Fig. P14.3-6. Write the expression (do not evaluate various integrals) for the P_{eM} of the optimum receiver, assuming all symbols to be equiprobable. Assume an AWGN channel.

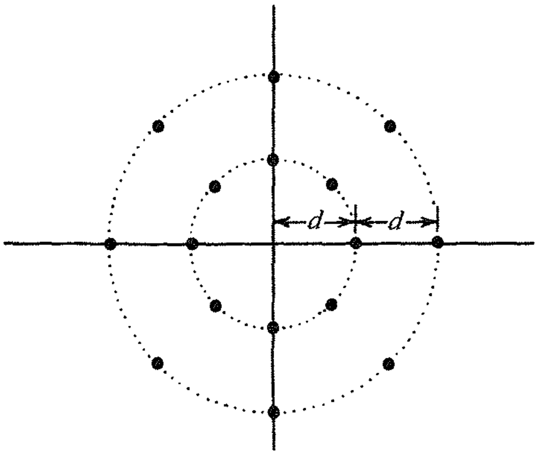


Figure P14.3-6

14.3-7 A five-signal configuration in a 2-dimensional space is shown in Fig. P14.3-7.

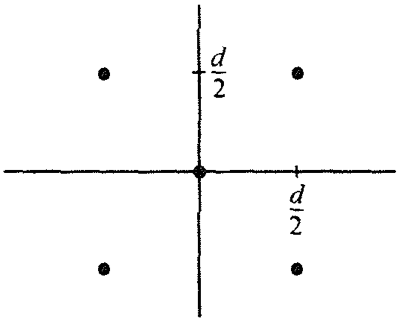


Figure P14.3-7

(**a**) Choose the appropriate $\varphi_1(t)$ and $\varphi_2(t)$ and sketch the waveforms of the five signals.

(**b**) In the signal space, sketch the optimum decision regions, assuming an AWGN channel.

(**c**) Determine the error probability P_{eM} as a function of $\bar{E}$ of the optimum receiver.

14.3-8 A 16-point QAM signal configuration is shown in Fig. P14.3-8. Assuming that all symbols are equiprobable, determine the error probability P_{eM} as a function of E_b of the optimum receiver for an AWGN channel.

Compare the performance of this scheme with that in Example 14.3.

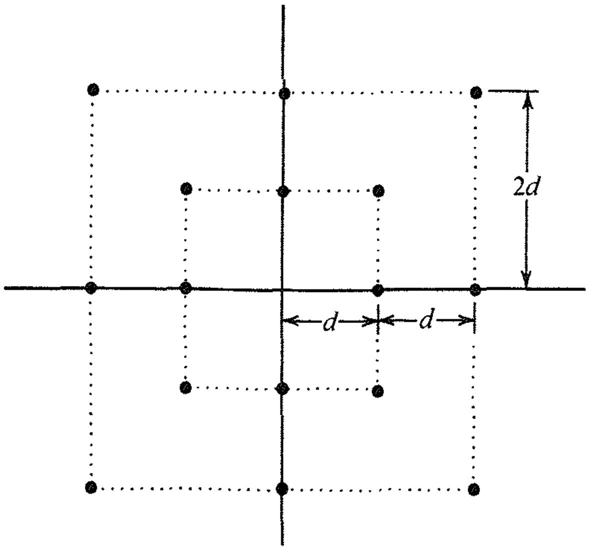


Figure P14.3-8

14.3-9 An orthogonal signal set is given by

$$s_k(t) = \sqrt{E}\,\varphi_k(t) \qquad k = 1, 2, \ldots, N$$

A biorthogonal signal set is formed from the orthogonal set by augmenting it with the negative of each signal. Thus, we add to the orthogonal set another set

$$s_{-k}(t) = -\sqrt{E}\,\varphi_k(t)$$

This gives $2N$ signals in an N-dimensional space. Assuming all signals to be equiprobable and an AWGN channel, obtain the error probability of the optimum receiver. How does the bandwidth of the biorthogonal set compare with that of the orthogonal set?

14.4-1 **(a)** What is the minimum-energy equivalent signal set of a binary on-off signal set?

(b) Using geometrical signal space concepts, explain why the binary on-off and the binary orthogonal sets have identical error probabilities and why the binary polar energy requirements are 3 dB lower than those of the on-off or the orthogonal set.

14.4-2 A source emits three equiprobable messages m_1, m_2, and m_3, encoded by signals $s_1(t)$, $s_2(t)$, and $s_3(t)$, respectively, where

$$\left.\begin{array}{rcl} s_1(t) & = & 20\sqrt{2}\sin\frac{2\pi}{T_M}t \\ s_2(t) & = & 10\sqrt{2}\cos\frac{2\pi}{T_M}t \\ s_3(t) & = & -10\sqrt{2}\cos\frac{2\pi}{T_M}t \end{array}\right\} \qquad T_M = \frac{1}{20}$$

Each of these signal durations is $0 \leq t \leq T_M$ and is zero outside this interval. The signals are transmitted over AWGN channels.

(a) Represent these signals in a signal space.

(b) Determine the decision regions.

(c) Obtain an equivalent minimum-energy signal set.

(d) Determine the optimum receiver.

14.4-3 A quaternary signaling scheme uses four waveforms,

$$s_1(t) = (\sqrt{3} - 1)\, \varphi_1(t)$$
$$s_2(t) = -2\varphi_1(t) + (\sqrt{3} - 1)\, \varphi_2(t)$$
$$s_3(t) = -(\sqrt{3} + 1)\, \varphi_1(t) - 2\varphi_2(t)$$
$$s_4(t) = -(\sqrt{3} + 1)\, \varphi(t)$$

where $\varphi_1(t)$ and $\varphi_2(t)$ are orthonormal basis signals. All the signals are equiprobable, and the channel noise is white gaussian with PSD $S_n(\omega) = 0.2$.

(a) Represent these signals in the signal space, and determine the optimum decision regions.

(b) Compute the error probability of the optimum receiver.

(c) Find the minimum-energy equivalent signal set.

14.4-4 A ternary signaling scheme ($M = 3$) uses the three waveforms $s_1(t)$, $s_2(t)$, and $s_3(t)$ shown in Fig. P14.4-4. The transmission rate is 500 symbols per second. All three messages are equiprobable, and the channel noise is white gaussian with PSD $S_n(\omega) = 10^{-5}$.

(a) Determine the error probability of the optimum receiver.

(b) Determine the minimum-energy signal set and sketch the waveforms.

(c) Compute the mean energies of the signal set in Fig. P14.4-4 and its minimum-energy equivalent set, found in part **(b)**.

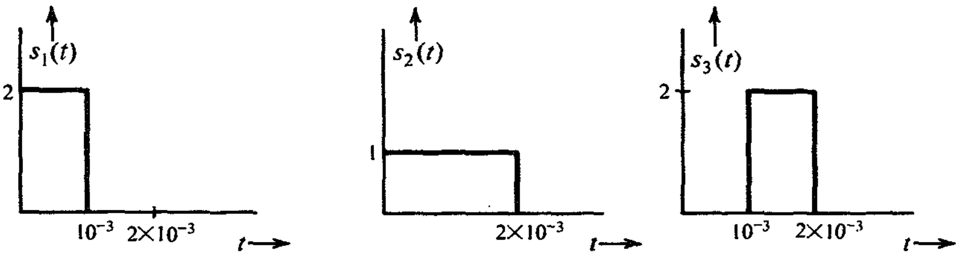


Figure P14.4-4

14.4-5 Repeat Prob. 14.4-4 if $P(m_1) = 0.25$, $P(m_2) = 0.5$, and $P(m_3) = 0.25$.

14.4-6 A binary signaling scheme uses the two waveforms $s_1(t)$ and $s_2(t)$ shown in Fig. P14.4-6. The signaling rate is 1000 pulses per second. Both signals are equally likely, and the channel noise is white gaussian with PSD $S_n(\omega) = 2.5 \times 10^{-6}$.

(a) Determine the minimum-energy equivalent signal set.

(b) Determine the error probability of the optimum receiver.

(c) Represent these signals as vectors using a suitable orthogonal signal space. *Hint*: Use Gram-Schmidt orthogonalization to determine the appropriate basis signals $\varphi_1(t)$ and $\varphi_2(t)$.

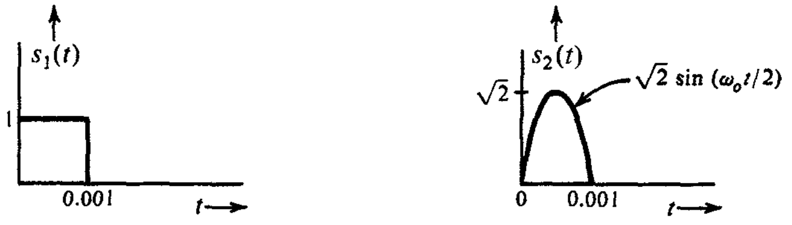


Figure P14.4-6

15

INTRODUCTION TO INFORMATION THEORY

In all the modes of communication discussed thus far, the communication is not error-free. We may be able to improve the accuracy in digital signals by reducing the error probability P_e. But it appears that as long as a channel noise exists, the communication cannot be error-free. For example, in all the digital systems discussed thus far, P_e varies as e^{-kE_b} asymptotically. By increasing E_b, the energy per bit, we can reduce P_e to any desired level. Now, the signal power is $S_i = E_b R_b$, where R_b is the bit rate. Hence, increasing E_b means either increasing the signal power S_i (for a given bit rate) or decreasing the bit rate R_b (for a given power), or both. Because of physical limitations, however, S_i cannot be increased beyond a certain limit. Hence, to reduce P_e further, we must reduce R_b, the rate of transmission of information digits. Thus, the price to be paid for reducing P_e is a reduction in the transmission rate. To make $P_e \to 0$, $R_b \to 0$. Hence, it appears that in the presence of channel noise it is impossible to achieve error-free communication. Thus thought communication engineers until the publication of Shannon's classical paper[1] in 1948. Shannon showed that for a given channel, as long as the rate of information digits per second to be transmitted is maintained within a certain limit (known as the channel capacity), it is possible to achieve error-free communication. That is, to attain $P_e \to 0$, it is not necessary to make $R_b \to 0$. Such a goal ($P_e \to 0$) can be attained by maintaining $R_b < C$, the channel capacity (per second). The gist of Shannon's paper is that the presence of random disturbance in a channel does not, by itself, set any limit on transmission accuracy. Instead, it sets a limit on the information rate for which arbitrarily small error probability ($P_e \to 0$) can be achieved.

We have been using the phrase "rate of information transmission" as if information could be measured. This is indeed so. We shall now discuss the information content of a message as understood by our "common sense" and also as it is understood in the "engineering sense." Surprisingly, both approaches yield the same measure of information in a message.

15.1 MEASURE OF INFORMATION

Common-Sense Measure of Information

Consider the following three hypothetical headlines in a morning paper:

1. Tomorow the sun will rise in the east.
2. United States invades Cuba.
3. Cuba invades the United States.

The reader will hardly notice the first headline. He or she will be very, very interested in the second. But what really catches the reader's fancy is the third one. This item will attract much more attention than the previous two headlines. From the viewpoint of "common sense," the first headline conveys hardly any information, the second conveys a large amount of information, and the third conveys yet a larger amount of information. If we look at the probabilities of occurrence of these three events, we find that the probability of occurrence of the first event is unity (a certain event), that of the second is very low (an event of small but finite probability), and that of the third is practically zero (an almost impossible event). If an event of low probability occurs, it causes greater surprise and, hence, conveys more information than the occurrence of an event of larger probability. Thus, the information is connected with the element of surprise, which is a result of uncertainty, or unexpectedness. The more unexpected the event, the greater the surprise, and hence the more information. The probability of occurrence of an event is a measure of its unexpectedness and, hence, is related to the information content. Thus, from the point of view of common sense, the amount of information received from a message is directly related to the uncertainty or inversely related to the probability of its occurrence. If P is the probability of occurrence of a message and I is the information gained from the message, it is evident from the preceding discussion that when $P \to 1, I \to 0$ and when $P \to 0, I \to \infty$, and, in general a smaller P gives a larger I. This suggests the following model:

$$I \sim \log \frac{1}{P} \tag{15.1}$$

Engineering Measure of Information

We now show that from an engineering point of view, the information content of a message is identical to that obtained on an intuitive basis [Eq. (15.1)]. What do we mean by an engineering point of view? An engineer is responsible for the efficient transmission of messages. For this service the engineer will charge a customer an amount proportional to the information to be transmitted. But in reality the engineer will charge the customer in proportion to the time required to transmit the message. In short, from an engineering point of view, the information in a message is proportional to the (minimum) time required to transmit the message. We shall now show that this concept of information also leads to Eq. (15.1). This implies that a message with higher probability can be transmitted in a shorter time than that required for a message with lower probability. This fact may be verified by the example of the transmission of alphabetic symbols in the English language using Morse code. This code is made up of various combinations of two symbols (such as a mark and a space, or pulses of height A and $-A$ volts). Each letter is represented by a certain combination of these symbols, called the **code word**, which has a certain length. Obviously, for efficient transmission, shorter code words are assigned to the letters e, t, a, and o, which occur more frequently. The longer code words are assigned to letters x, k, q, and z, which occur less frequently. Each letter may be considered as a message. It is obvious that the letters that occur more frequently (with higher probability of occurrence) need a shorter time to transmit (shorter code words) as compared to those with smaller probability of occurrence. We shall now show that on the average, the time required to transmit a symbol (or a message) with probability of occurrence P is indeed proportional to $\log(1/P)$.

For the sake of simplicity, let us begin with the case of binary messages m_1 and m_2, which are equally likely to occur. We may use binary digits to encode these messages. Messages m_1 and m_2 may be represented by digits **0** and **1**, respectively. Clearly, we must have a minimum of one binary digit (which can assume two values) to represent each of the two equally likely messages. Next, consider the case of the four equiprobable messages m_1, m_2, m_3, and m_4. If these messages are encoded in binary form, we need a minimum of two binary digits per message. Each binary digit can assume two values. Hence, a combination of two binary digits can form the four code words **00**, **01**, **10**, **11**, which can be assigned to the four equiprobable messages m_1, m_2, m_3, and m_4, respectively. It is clear that each of these four messages takes twice as much transmission time as that required by each of the two equiprobable messages and, hence, contains twice as much information. Similarly, we can encode any one of eight equiprobable messages with a minimum of three binary digits. This is because three binary digits form eight distinct code words, which can be assigned to each of the eight messages. It can be seen that, in general, we need $\log_2 n$ binary digits to encode each of n equiprobable messages.* Because all the messages are equiprobable, P, the probability of any one message occurring, is $1/n$. Hence, each message (with probability P) needs $\log_2(1/P)$ binary digits for encoding. Thus, from the engineering viewpoint, the information I contained in a message with probability of occurrence P is proportional to $\log_2(1/P)$,

$$I = k \log_2 \frac{1}{P} \tag{15.2}$$

where k is a constant to be determined. Once again, we come to the conclusion (from the engineering viewpoint) that the information content of a message is proportional to the logarithm of the reciprocal of the probability of the message.

We shall now define the information conveyed by a message according to Eq. (15.2). The constant of proportionality is taken as unity for convenience, and the information is then in terms of binary units, abbreviated **bit** (**bin**ary uni**t**),

$$I = \log_2 \frac{1}{P} \quad \text{bits} \tag{15.3}$$

According to this definition, the information I in a message can be interpreted as the minimum number of binary digits required to encode the message. This is given by $\log_2(1/P)$, where P is the probability of occurrence of the message. Although here we have shown this result for the special case of equiprobable messages, we shall show in the next section that this is true for nonequiprobable messages also.

Next, we shall consider the case of r-ary digits instead of binary digits for encoding. Each of the r-ary digits can assume r values $(\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{r-1})$. Each of n messages (encoded by r-ary digits) can then be transmitted by a particular sequence of r-ary signals. Because each r-ary digit can assume r values, k r-ary digits can form a maximum of r^k distinct code words. Hence, to encode each of the n equiprobable messages, we need a minimum of $k = \log_r n$ r-ary digits.[*] But $n = 1/P$, where P is the probability of occurrence of each message. Obviously, we need a minimum of $\log_r(1/P)$ r-ary digits. The information I per message is therefore

---

* Here we are assuming that the number n is such that $\log_2 n$ is an integer. Later on we shall observe that this restriction is not necessary.

$$I = \log_r \frac{1}{P} \quad r\text{-ary units} \tag{15.4}$$

From Eqs. (15.3) and (15.4) it is evident that

$$I = \log_2 \frac{1}{P} \quad \text{bits} = \log_r \frac{1}{P} \quad r\text{-ary units}$$

Hence,[†]

$$1\ r\text{-ary unit} = \log_2 r \text{ bits} \tag{15.5}$$

A Note on the Unit of Information: Although it is tempting to use the r-ary unit as a general unit of information, the binary unit bit ($r = 2$) is commonly used in the literature. There is, of course, no loss of generality in using $r = 2$. These units can always be converted into any other units by using Eq. (15.5). Henceforth, we shall always use the binary unit (bit) for information, unless otherwise stated. The bases of the logarithmic functions will be omitted, but will be understood to be 2.

Average Information per Message: Entropy of a Source

Consider a memoriless source m emitting messages $m_1, m_2, \ldots, m_n$ with probabilities $P_1, P_2, \ldots, P_n$, respectively ($P_1 + P_2 + \cdots + P_n = 1$). A **memoriless source** implies that each message emitted is independent of the previous message(s). By the definition in Eq. (15.3) [or Eq. (15.4)], the information content of message m_i is I_i, given by

$$I_i = \log \frac{1}{P_i} \quad \text{bits} \tag{15.6}$$

The probability of occurrence of m_i is P_i. Hence, the mean, or average, information per message emitted by the source is given by $\sum_{i=1}^{n} P_i I_i$ bits. The average information per message of a source m is called its **entropy**, denoted by $H(\text{m})$. Hence,

$$H(\text{m}) = \sum_{i=1}^{n} P_i I_i \quad \text{bits}$$

---

[*] Here again we are assuming that n is such that $\log_r n$ is an integer. As we shall see later, this restriction is not necessary.

[†] In general,

$$1\ r\text{-ary unit} = \log_s r\ s\text{-ary units}$$

The 10-ary unit of information is called the **hartley** in honor of R. V. L. Hartley,[2] who was one of the pioneers (along with Nyquist[3] and Carson) in the area of information transmission in the 1920s. The rigorous mathematical foundations of information theory, however, were established by E. Shannon[1] in 1948:

$$1 \text{ hartley} = \log_2 10 = 3.32 \text{ bits}$$

Sometimes the unit **nat** is used:

$$1 \text{ nat} = \log_2 e = 1.44 \text{ bits}$$

$$= \sum_{i=1}^{n} P_i \log \frac{1}{P_i} \quad \text{bits} \tag{15.7a}$$

$$= -\sum_{i=1}^{n} P_i \log P_i \quad \text{bits} \tag{15.7b}$$

The entropy of a source is a function of the message probabilities. It is interesting to find the message probability distribution that yields the maximum entropy. Because the entropy is a measure of uncertainty, the probability distribution that generates the maximum uncertainty will have the maximum entropy. On qualitative grounds, one expects entropy to be maximum when all the messages are equiprobable. We shall now show that this is indeed true.

Because $H(\mathrm{m})$ is a function of $P_1, P_2, \ldots, P_n$, the maximum value of $H(\mathrm{m})$ is found from the equation $dH(\mathrm{m})/dP_i = 0$ for $i = 1, 2, \ldots, n$, with the constraint that

$$P_n = 1 - (P_1 + P_2 + \cdots + P_{n-1}) \tag{15.8}$$

Because

$$H(\mathrm{m}) = -\sum_{i=1}^{n} P_i \log P_i \tag{15.9}$$

we need consider only the terms $-P_i \log P_i$ and $-P_n \log P_n$ [because P_n is a function of P_i, as seen from Eq. (15.8)]. Hence,

$$\begin{aligned}
\frac{dH(\mathrm{m})}{dP_i} &= \frac{d}{dP_i}(-P_i \log P_i - P_n \log P_n) \\
&= -P_i \left(\frac{1}{P_i}\right) \log e - \log P_i + P_n \left(\frac{1}{P_n}\right) \log e + \log P_n \\
&= \log \frac{P_n}{P_i}
\end{aligned}$$

which is zero if $P_i = P_n$. Because this is true for all i, we have

$$P_1 = P_2 = \cdots = P_n = \frac{1}{n} \tag{15.10}$$

To show that Eq. (15.10) yields $[H(\mathrm{m})]_{\max}$ and not $[H(\mathrm{m})]_{\min}$, we note that when $P_1 = 1$ and $P_2 = P_3 = \cdots = P_n = 0$, $H(\mathrm{m}) = 0$, whereas the probabilities in Eq. (15.10) yield

$$\begin{aligned}
H(\mathrm{m}) &= -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} \\
&= \log n
\end{aligned} \tag{15.11}$$

The Intuitive (Common-Sense) and the Engineering Interpretation of Entropy: Earlier we observed that both the intuitive and the engineering viewpoints lead to the same definition of the information associated with a message. The conceptual bases, however, are entirely different for the two points of view. Consequently, we have two physical interpretations of information. According to the engineering point of view, the information content of any message is equal to the minimum number of digits required to encode the message, and, therefore, the entropy $H(\mathrm{m})$ is equal to the minimum number of digits per message required, on the average, for encoding. From the intuitive standpoint, on the other hand, information is thought of as being synonymous with the amount of surprise, or uncertainty, associated with the event (or message). A smaller probability of occurrence implies more uncertainty about the event. Uncertainty is, of course, associated with surprise. Hence intuitively, the information associated with a message is a measure of the uncertainty (unexpectedness) of the message. Therefore, $\log(1/P_i)$ is a measure of the uncertainty of the message m_i, and $\sum_{i=1}^{n} P_i \log(1/P_i)$ is the average uncertainty (per message) of the source that generates messages $m_1, m_2, \ldots, m_n$ with probabilities $P_1, P_2, \ldots, P_n$. Both these interpretations prove useful in the qualitative understanding of the mathematical definitions and results in information theory. Entropy may also be viewed as a function associated with a random variable m that assumes values $m_1, m_2, \ldots, m_n$ with probabilities $P(m_1), P(m_2), \ldots, P(m_n)$:

$$
\begin{aligned}
H(\mathrm{m}) &= \sum_{i=1}^{n} P(m_i) \log \frac{1}{P(m_i)} \\
&= \sum_{i=1}^{n} P_i \log \frac{1}{P_i}
\end{aligned}
$$

Thus, we can associate an entropy with every discrete random variable.

If the source is not memoriless (i.e., a message emitted at any time is not independent of the previous messages emitted), then the source entropy will be less than $H(\mathrm{m})$ in Eq. (15.9). This is because the dependence of a message on previous messages reduces its uncertainty.

15.2 SOURCE ENCODING

The minimum number of binary digits required to encode a message was shown to be equal to the source entropy $\log(1/P)$ if all the messages of the source are equiprobable (each message probability is P). We shall now generalize this result to the case of nonequiprobable messages. We shall show that the average number of binary digits per message required for encoding is given by $H(\mathrm{m})$ (in bits) for an arbitrary probability distribution of messages.

Let a source m emit messages $m_1, m_2, \ldots, m_n$ with probabilities $P_1, P_2, \ldots, P_n$, respectively. Consider a sequence of N messages with $N \to \infty$. Let k_i be the number of times message m_i occurs in this sequence. Then according to the relative frequency interpretation (or law of large numbers),

$$
\lim_{N\to\infty} \frac{k_i}{N} = P_i
$$

Thus, the message m_i occurs NP_i times in a sequence of N messages (provided $N \to \infty$). Therefore, in a typical sequence of N messages, m_1 will occur NP_1 times, m_2 will occur NP_2 times, ..., m_n will occur NP_n times. All other compositions are extremely unlikely to occur ($P \to 0$). Thus, any typical sequence (where $N \to \infty$) has the same proportion of the n messages, although in general the order will be different. We shall assume a zero-memory source, that is, the message is emitted from the source independently of the previous messages.

Consider now a typical sequence S_N of N messages from the source. Because the n messages (of probability $P_1, P_2, \ldots, P_n$) occur $NP_1, NP_2, \ldots, NP_n$ times, and because each message is independent, the probability of occurrence of a typical sequence S_N is given by

$$P(S_N) = (P_1)^{NP_1}(P_2)^{NP_2}\cdots(P_n)^{NP_n} \tag{15.12}$$

Because all possible sequences of N messages from this source have the same composition, all the sequences (of N messages) are equiprobable, with probability $P(S_N)$. We can consider these long sequences as new messages (which are now equiprobable). To encode one such sequence we need L_N binary digits, where

$$L_N = \log\left[\frac{1}{P(S_N)}\right] \quad \text{binary digits} \tag{15.13}$$

Substituting Eq. (15.12) into Eq. (15.13), we obtain

$$\begin{aligned} L_N &= N\sum_{i=1}^{n} P_i \log\frac{1}{P_i} \\ &= NH(\mathrm{m}) \quad \text{binary digits} \end{aligned}$$

Note that L_N is the length (number of binary digits) of the code word required to encode N messages in sequence. Hence, L, the average number of digits required per message, is L_N/N and is given by

$$L = \frac{L_N}{N} = H(\mathrm{m}) \quad \text{binary digits} \tag{15.14}$$

This is the desired result, which states that it is possible to encode the messages emitted by a source using, on the average, $H(\mathrm{m})$ number of binary digits per message, where $H(\mathrm{m})$ is the entropy of the source (in bits). Although it does not prove that, on the average, this is the minimum number of digits required, one can show that $H(\mathrm{m})$ is indeed the minimum. It is not possible to find any uniquely decodable code whose average length is less than $H(\mathrm{m})$.[4,5]

Compact Codes

The source encoding theorem says that to encode a source with entropy $H(\mathrm{m})$, we need, on the average, a minimum of $H(\mathrm{m})$ binary digits per message, or $H_r(\mathrm{m})$ r-ary digits per message, where $H_r(\mathrm{m})$ is the entropy in Eq. (15.9) computed with r as the base of the logarithm. The number of digits in the code word is the **length** of the code word. Thus, the average word length of an optimum code is $H(\mathrm{m})$. Unfortunately, to attain this length, in general, we have to encode a sequence of N messages ($N \to \infty$) at a time. If we wish to encode each message directly without using longer sequences, then, in general, the average length of the code word per message will be greater than $H(\mathrm{m})$. In practice, it is not desirable to use long sequences, as they cause transmission delay and add to equipment complexity. Hence, it is preferable to encode messages directly, even if the price has to be paid in terms of increased word length. In most cases, the price turns out to be small. The following is a procedure, given without proof, for finding the optimum source code, called the Huffman code. The proof that this code is optimum can be found elsewhere.[4–6]

We shall illustrate the procedure with an example using a binary code. We first arrange the messages in the order of descending probability, as shown in Table 15.1. Here we have

Table 15.1

Original Source		Reduced Sources			
Messages	**Probabilities**	S_1	S_2	S_3	S_4
m_1	0.30	0.30	0.30	0.43	0.57
m_2	0.25	0.25	0.27	0.30	0.43
m_3	0.15	0.18	0.25	0.27	
m_4	0.12	0.15	0.18		
m_5	0.10	0.12			
m_6	0.08				

six messages with probabilities 0.30, 0.25, 0.15, 0.12, 0.10, and 0.08, respectively. We now combine the last two messages into one message with probability $P_5 + P_6 = 0.18$. This leaves five messages with probabilities, 0.30, 0.25, 0.15, 0.12, and 0.18. These messages are now rearranged in the second column in the order of descending probability. We repeat this procedure by combining the last two messages in the second column and rearranging them in the order of descending probability. This is done until the number of messages is reduced to 2. These two (reduced) messages are now assigned **0** and **1** as their first digits in the code sequence. We now go back and assign the numbers **0** and **1** to the second digit for the two messages that were combined in the previous step. We keep regressing this way until the first column is reached. The code finally obtained (for the first column) can be shown to be optimum. The complete procedure is shown in Tables 15.1 and 15.2.

The optimum (Huffman) code obtained this way is also called a **compact code**. The average length of the compact code in the present case is given by

$$L = \sum_{i=1}^{n} P_i L_i = 0.3(2) + 0.25(2) + 0.15(3) + 0.12(3) + 0.1(3) + 0.08(3)$$

$$= 2.45 \text{ binary digits}$$

The entropy $H(\mathrm{m})$ of the source is given by

$$H(\mathrm{m}) = \sum_{i=1}^{n} P_i \log_2 \frac{1}{P_i}$$

$$= 2.418 \text{ bits}$$

Table 15.2

Original Source			Reduced Sources							
Messages	**Probabilities**	**Code**	S_1		S_2		S_3		S_4	
m_1	0.30	**00**	0.30	**00**	0.30	**00**	0.43	**1**	0.57	**0**
m_2	0.25	**10**	0.25	**10**	0.27	**01**	0.30	**00**	0.43	**1**
m_3	0.15	**010**	0.18	**11**	0.25	**10**	0.27	**01**		
m_4	0.12	**011**	0.15	**010**	0.18	**11**				
m_5	0.10	**110**	0.12	**011**						
m_6	0.08	**111**								

Hence, the minimum possible length (attained by an infinitely long sequence of messages) is equal to 2.418 binary digits. Using direct coding (the Huffman code), it is possible to attain an average length of 2.45 bits in the example given. This is a close approximation of the optimum performance attainable. Thus, little is gained by complex coding using long sequences of messages in this case.

The merit of any code is measured by its average length in comparison to $H(\mathrm{m})$ (the average minimum length). We define the **code efficiency** η as

$$\eta = \frac{H(\mathrm{m})}{L}$$

where L is the average length of the code. In our present example,

$$\begin{aligned}\eta &= \frac{2.418}{2.45} \\ &= 0.976\end{aligned}$$

The **redundancy** γ is defined as

$$\begin{aligned}\gamma &= 1 - \eta \\ &= 0.024\end{aligned}$$

Huffman code is uniquely decodable. If we receive a sequence of Huffman-coded messages, it can be decoded only one way, that is, without ambiguity. For instance, if the source in this example were to emit the following message sequence: $m_1m_5m_2m_1m_4m_3m_6\ldots$, it would be encoded as **001101000011010111**.... The reader may verify that this message sequence can be decoded only one way, viz, $m_1m_5m_2m_1m_4m_3m_6\ldots$, even if there is no demarcation between individual messages.

A similar procedure is used to find a compact r-ary code. In this case we arrange the messages in descending order of probability, combine the last r messages into one message, and rearrange the new set (reduced set) in the order of descending probability. We repeat the procedure until the final set reduces to r messages. Each of these messages is now assigned one of the r numbers $\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{r} - \mathbf{1}$. We now regress in exactly the same way as in the binary case until each of the original messages is assigned a code.

For an r-ary code, we will have exactly r messages left in the last reduced set if, and only if, the total number of original messages is equal to $r + k(r - 1)$, where k is an integer. This is because each reduction decreases the number of messages by $r - 1$. Hence, if there is a total of k reductions, the total number of original messages must be $r + k(r - 1)$. In case the original messages do not satisfy this condition, we must add some dummy messages with zero probability of occurrence until this condition is fulfilled. As an example, if $r = 4$ and the number of messages n is 6, then we must add one dummy message with zero probability of occurrence to make the total number of messages 7, that is, $[4 + 1(4 - 1)]$, and proceed as usual. The procedure is illustrated in Example 15.1.

EXAMPLE 15.1 A zero-memory source emits six messages with probabilities 0.3, 0.25, 0.15, 0.12, 0.1, and 0.08. Find the 4-ary (quaternary) Huffman code. Determine its average word length, the efficiency, and the redundancy.

In this case, we need to add one dummy message to satisfy the required condition of $r + k(r - 1)$ messages and proceed as usual. The Huffman code is found in Table 15.3. The length L of this code is

$$\begin{aligned} L &= 0.3(1) + 0.25(1) + 0.15(1) + 0.12(2) + 0.1(2) + 0.08(2) + 0(2) \\ &= 1.3 \quad \text{4-ary digits} \end{aligned}$$

Also,

$$\begin{aligned} H_4(\mathrm{m}) &= -\sum_{i=1}^{6} P_i \log_4 P_i \\ &= 1.209 \quad \text{4-ary units} \end{aligned}$$

The code efficiency η is given by

$$\eta = \frac{1.209}{1.3} = 0.93$$

The redundancy $\gamma = 1 - \eta = 0.07$.

Table 15.3

Original Source				
Messages	**Probabilities**	**Code**	**Reduced Sources**	
m_1	0.30	**0**	0.30	**0**
m_2	0.25	**2**	0.30	**1**
m_3	0.15	**3**	0.25	**2**
m_4	0.12	**10**	0.15	**3**
m_5	0.10	**11**		
m_6	0.08	**12**		
m_7	0.00	**13**		

To achieve code efficiency $\eta \to 1$, we need $N \to \infty$. The Huffman code uses $N = 1$, but its efficiency is, in general, less than 1. A compromise exists between these two extremes of $N = 1$ and $N = \infty$. We can use $N = 2$ or 3. In most cases, the use of $N = 2$ or 3 can yield an efficiency close to 1, as the following example shows.

EXAMPLE 15.2 A zero-memory source emits messages m_1 and m_2 with probabilities 0.8 and 0.2, respectively. Find the optimum (Huffman) binary code for this source as well as for its **second-** and **third-order extensions** (that is, for $N = 2$ and 3). Determine the code efficiencies in each case.

The Huffman code for the source is simply **0** and **1**, giving $L = 1$, and

$$\begin{aligned} H(\mathrm{m}) &= -(0.8 \log 0.8 + 0.2 \log 0.2) \\ &= 0.72 \quad \text{bit} \end{aligned}$$

Hence,

$$\eta = 0.72$$

For the second-order extension of the source ($N = 2$), there are four possible composite messages, m_1m_1, m_1m_2, m_2m_1, and m_2m_2, with probabilities 0.64, 0.16, 0.16, and 0.04, respectively. The Huffman code is obtained in Table 15.4.

Table 15.4

Original Source						
Messages	**Probabilities**	**Code**	**Reduced Source**			
m_1m_1	0.64	**0**	0.64	**0**	0.64	**0**
m_1m_2	0.16	**11**	0.20	**10**	0.36	**1**
m_2m_1	0.16	**100**	0.16	**11**		
m_2m_2	0.04	**101**				

In this case the average word length L' is

$$\begin{aligned} L' &= 0.64(1) + 0.16(2) + 0.16(3) + 0.04(3) \\ &= 1.56 \end{aligned}$$

This is the word length for two messages of the original source. Hence L, the word length per message, is

$$L = \frac{L'}{2} = 0.78$$

and

$$\eta = \frac{0.72}{0.78} = 0.923$$

If we proceed with $N = 3$ (the third-order extension of the source), we have eight possible messages, and following the Huffman procedure, we find the code as shown in Table 15.5.

Table 15.5

Messages	Probabilities	Code
$m_1m_1m_1$	0.512	**0**
$m_1m_1m_2$	0.128	**100**
$m_1m_2m_1$	0.128	**101**
$m_2m_1m_1$	0.128	**110**
$m_1m_2m_2$	0.032	**11100**
$m_2m_1m_2$	0.032	**11101**
$m_2m_2m_1$	0.032	**11110**
$m_2m_2m_2$	0.008	**11111**

The word length L'' is

$$\begin{aligned} L'' &= (0.512)1 + (0.128 + 0.128 + 0.128)3 \\ &\quad + (0.032 + 0.032 + 0.032)5 + (0.008)5 \\ &= 2.184 \end{aligned}$$

Then,

$$L = \frac{L''}{3} = 0.728$$

and

$$\eta = \frac{0.72}{0.728} = 0.989$$

15.3 ERROR-FREE COMMUNICATION OVER A NOISY CHANNEL

As seen in the previous section, messages of a source with entropy $H(\mathrm{m})$ can be encoded by using an average of $H(\mathrm{m})$ digits per message. This encoding has zero redundancy. Hence, if we transmit these coded messages over a noisy channel, some of the information will be received erroneously. There is absolutely no possibility of error-free communication over a noisy channel when messages are encoded with zero redundancy. The use of redundancy, in general, helps combat noise. This can be seen from a simple example of a **single parity-check code**, in which an extra binary digit is added to each code word to ensure that the total number of **1**'s in the resulting code word is always even (or odd). If a single error occurs in the received code word, the parity is violated, and the receiver requests retransmission. This is a rather simple example to demonstrate the utility of redundancy. More complex coding procedures, which can correct up to n digits, will be discussed in the next chapter.

The addition of an extra digit increases the average word length to $H(\mathrm{m}) + 1$, giving $\eta = H(\mathrm{m})/[H(\mathrm{m}) + 1]$, and the redundancy is $1 - \eta = 1/[H(\mathrm{m}) + 1]$. Thus, the addition of an extra check digit increases redundancy, but it also helps combat noise. Immunity against channel noise can be increased by increasing the redundancy. Shannon has shown that it is possible to achieve error-free communication by adding sufficient redundancy. For example, if we have a **binary symmetric channel (BSC)** with an error probability P_e, then for error-free communication over this channel, messages from a source with entropy $H(\mathrm{m})$ must be encoded by binary codes with a word length of at least $H(\mathrm{m})/C_s$, where (see Sec. 15.4)

$$C_s = 1 - \left[P_e \log \frac{1}{P_e} + (1 - P_e) \log \frac{1}{1 - P_e}\right] \tag{15.15}$$

The parameter C_s ($C_s < 1$) is called the **channel capacity**.

The efficiency of these codes is never greater than C_s. If a certain binary channel has $C_s = 0.4$, a code that can achieve error-free communication must have at least 2.5 $H(\mathrm{m})$ binary digits per message, which is two-and-one-half times as many digits as required for coding without redundancy. This means there are 1.5 $H(\mathrm{m})$ redundant digits per message. Thus, on the average, for every 2.5 digits transmitted, one digit is the information digit and 1.5 digits are redundant, or check, digits, giving a redundancy of $1 - C_s = 0.6$.

As discussed in the beginning of this chapter, P_e, the error probability of binary signaling, varies as e^{-kE_b} and, hence, to make $P_e \to 0$, either $S_i \to \infty$ or $R_b \to 0$. Because S_i must be finite, $P_e \to 0$ only if $R_b \to 0$. But Shannon's results state that it is really not necessary to let $R_b \to 0$ for error-free communication. All that is required is to hold R_b below C, the channel capacity per second ($C = 2BC_s$). Where is the discrepancy? To answer this question let us investigate carefully the role of redundancy in error-free communication. Although the discussion here is with reference to a binary scheme, it is quite general and can be extended to the M-ary case.

Consider a simple method of reducing P_e by repeating a given digit an odd number of times. For example, we can transmit **0** and **1** as **000** and **111**. The receiver uses the majority rule to make the decision; that is, if at least two out of three digits are **1**, the decision is **1**, and if at least two out of three digits are **0**, the decision is **0**. Thus, even if one out of three digits is in error, the information is received error-free. This scheme will fail if two out of three digits are in error. In order to correct two errors, we need five repetitions. In any case, repetitions cause redundancy but improve P_e (see Example 10.8). It will be instructive to understand this situation from a graphic point of view. Consider the case of three repetitions. We can show all eight possible sequences of three binary digits graphically as the vertices of a cube (Fig. 15.1). It is convenient to map binary sequences as shown in Fig. 15.1 and to talk in terms of what is called the **Hamming distance** between binary sequences. If two binary sequences of the same length differ in j places (j digits), then the Hamming distance between the sequences is considered to be j. Thus, the Hamming distance between **000** and **010** (or **001** and **101**) is 1, and between **000** and **111** it is 3. In the case of three repetitions, we transmit binary **1** by **111** and binary **0** by **000**. The Hamming distance between these two sequences is 3. Observe that of the eight possible vertices, we are occupying only two (**000** and **111**) for transmitted messages. At the receiver, however, because of channel noise, we are liable to receive any one of the eight sequences. The majority decision rule can be interpreted as a rule that decides in favor of that message (**000** or **111**), which is at the closest Hamming distance from the received sequence. Sequences **000**, **001**, **010**, and **100** are within 1 unit of Hamming distance from **000** but are at least 2 units away from **111**. Hence, when we receive any one of these four sequences, our decision is binary **0**. Similarly, when any one of the sequences **110**, **111**, **011**, or **101** is received, the decision is binary **1**.

We can now see why the error probability is reduced in this scheme. Of the eight possible vertices, we have used only two, which are separated by 3 Hamming units. If we draw a Hamming sphere of unit radius around each of these two vertices (**000** and **111**), the two Hamming spheres* are nonoverlapping. The channel noise can cause a distance between the received sequence and the transmitted sequence, and as long as this distance is equal to or less than 1 unit, we can still detect the message without error. In a similar way, the case of five repetitions can be represented by a hypercube of five dimensions. The transmitted sequences

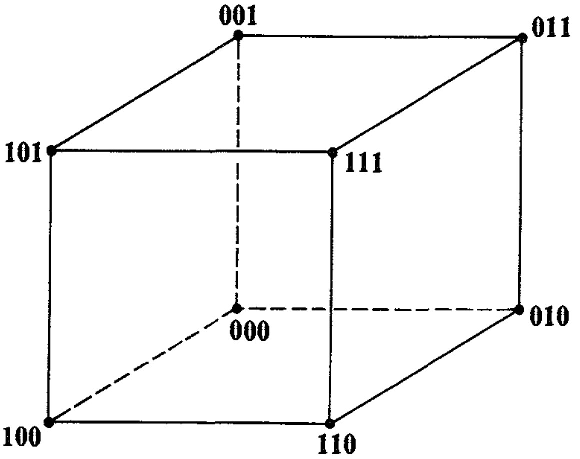


Figure 15.1 3-dimensional cube in Hamming space.

---

* Note that the Hamming sphere is not a true geometrical hypersphere because the Hamming distance is not a true geometrical distance (e.g., sequences **001**, **010**, and **100** lie on a Hamming sphere centered at **111** and having a radius 2).

00000 and **11111** occupy two vertices separated by five units, and the Hamming spheres of 2-unit radius drawn around each of these two vertices would be nonoverlapping. In this case, even if channel noise causes two errors, we can still detect the message correctly. Hence, the reason for the reduction in error probability is that we have not used all the available vertices for messages. Had we occupied all the available vertices for messages (as is the case without redundancy, or repetition), then if channel noise caused an error (even one), the received sequence would occupy a vertex assigned to another transmitted sequence, and we are certain to make a wrong decision. Precisely because we have left the neighboring vertices of the transmitted sequence unoccupied, are we able to detect the sequence correctly, despite channel errors (within a certain limit). The smaller the fraction of vertices used, the smaller the error probability. It should also be remembered that redundancy (or repetition) is what makes it possible to have unoccupied vertices.

If we continue to increase n, the number of repetitions, we will reduce P_e, but we will also reduce R_b by the factor n. But no matter how large we make n, the error probability never becomes zero. The trouble with this scheme is that it is inefficient because we are adding redundant (or check) digits to each information digit. To give an analogy, redundant (or check) digits are like guards protecting the information digit. To hire guards for each information digit is somewhat similar to a case of families living on a certain street hit by several burglaries. Each family panics and hires a guard. This is obviously expensive and inefficient. A better solution would be for all the families on the street to hire one guard and share the expenses. One guard can check on all the houses on the street, assuming a street of reasonable size. If the street is too long, it might be necessary to hire more than one guard. But it is certainly not necessary to hire one guard per house. In using repetitions, we had a similar situation. Redundant (or repeated) digits were used to check on only one transmitted digit. Using the clue from the preceding analogy, it might be more efficient if we used redundant digits not to check (guard) any one individual transmitted digit but, rather, a block of digits. Herein lies the key to our problem. Let us consider a group of information digits over a certain time interval of T seconds, and let us add some redundant digits to check on all these digits.

Suppose we need to transmit α binary information digits per second. Then over a period of T seconds, we have a block of αT binary information digits. If to this block of information digits we add $(\beta - \alpha)T$ check digits ($\beta - \alpha$ check digits, or redundant digits, per second), then we need to transmit βT ($\beta > \alpha$) digits for every αT information digits. Therefore over a T-second interval, we have

αT	information digits
βT	total transmitted digits ($\beta > \alpha$)
$(\beta - \alpha)T$	check digits

Thus, instead of transmitting one binary digit every $1/\alpha$ seconds, we let αT digits accumulate over T seconds. Now consider this as a message to be transmitted. There are a total of $2^{\alpha T}$ such supermessages. Thus, every T seconds we need to transmit one of the $2^{\alpha T}$ possible supermessages. These supermessages are transmitted by a sequence of βT binary digits. There are in all $2^{\beta T}$ possible sequences of βT binary digits, and they can be represented as vertices of a βT-dimensional hypercube. Because we have only $2^{\alpha T}$ messages to be transmitted whereas $2^{\beta T}$ vertices are available, we occupy only a $2^{-(\beta-\alpha)T}$ fraction of the vertices of the βT-

dimensional hypercube. Observe that we have reduced the transmission rate by a factor of α/β. This rate-reduction factor α/β is independent of T. The fraction of the vertices occupied (occupancy factor) by transmitted messages is $2^{-(\beta-\alpha)T}$ and can be made as small as possible simply by increasing T. In the limit as $T \to \infty$, the occupancy factor approaches 0. This will make the error probability go to 0, and we have the possibility of error-free communication.

One important question, however, remains to be answered. What must be the rate reduction ratio α/β for this dream to come true? To answer this question, we observe that increasing T increases the length of the transmitted sequence (βT digits). If P_e is the digit error probability, then it can be seen from the relative frequency definition (or the law of large numbers) that as $T \to \infty$, the total number of digits in error in a sequence of βT digits ($\beta T \to \infty$) is exactly $\beta T P_e$. Hence, the received sequences will be at a Hamming distance of $\beta T P_e$ from the transmitted sequences. Therefore, for error-free communication, we must leave all the vertices unoccupied within spheres of radius $\beta T P_e$ drawn around each of the $2^{\alpha T}$ occupied vertices. In short, we must be able to pack $2^{\alpha T}$ nonoverlapping spheres, each of radius $\beta T P_e$, into the Hamming space of βT dimensions. This means that for a given β, α cannot be increased beyond some limit without causing overlap in the spheres and the consequent failure of the scheme. Shannon's theorem states that for this scheme to work, α/β must be less than the constant (channel capacity) C_s, which is a function of the channel noise and the signal power:

$$\frac{\alpha}{\beta} < C_s \tag{15.16}$$

It must be remembered that such perfect, error-free communication is not practical. In this system we accumulate the information digits for T seconds before encoding them, and because $T \to \infty$, for error-free communication we must wait until eternity before we start encoding. Hence, there will be an infinite delay at the transmitter and an additional delay of the same amount at the receiver. Second, the equipment needed for the storage, encoding, and decoding sequence of infinite digits would be monstrous. Needless to say that in practice the dream of error-free communication cannot be achieved. Then what is the use of Shannon's result? For one thing, it indicates the upper limit on the rate of error-free communication that can be achieved on a channel. This in itself is monumental. Second, it indicates the way to reduce the error probability with only a small reduction in the rate of transmission of information digits. We can therefore seek a compromise between error-free communication with infinite delay and virtually error-free communication with a finite delay.

15.4 CHANNEL CAPACITY OF A DISCRETE MEMORILESS CHANNEL

In this section, discrete memoriless channels will be considered. Let a source emit symbols $x_1, x_2, \ldots, x_r$. The receiver receives symbols $y_1, y_2, \ldots, y_s$. The set of symbols $\{y_k\}$ may or may not be identical to the set $\{x_k\}$, depending on the nature of the receiver. If we use the types of receivers discussed in Chapter 14, the set of received symbols will be the same as the set transmitted. This is because the optimum receiver, upon receiving a signal, decides which of the r symbols $x_1, x_2, \ldots, x_r$ has been transmitted. Here we shall be more general and shall not constrain the set $\{y_k\}$ to be identical to the set $\{x_k\}$.

If the channel is noiseless, then the reception of some symbol y_j uniquely determines the message transmitted. Because of noise, however, there is a certain amount of uncertainty regarding the transmitted symbol when y_j is received. If $P(x_i|y_j)$ represents the conditional probabilities that x_i was transmitted when y_j is received, then there is an uncertainty of $\log[1/P(x_i|y_j)]$ about x_i when y_j is received. When this uncertainty is averaged over all x_i and y_j, we obtain $H(\mathrm{x}|\mathrm{y})$, which is the average uncertainty about a transmitted symbol when a symbol is received. Thus,

$$H(\mathrm{x}|\mathrm{y}) = \sum_i \sum_j P(x_i, y_j) \log \frac{1}{P(x_i|y_j)} \quad \text{bits per symbol} \tag{15.17}$$

If the channel were noiseless, the uncertainty would be zero.* Obviously, this uncertainty, $H(\mathrm{x}|\mathrm{y})$, is caused by channel noise. Hence, it is the average loss of information about a transmitted symbol when a symbol is received. We call $H(\mathrm{x}|\mathrm{y})$ the **equivocation** of x with respect to y.

Note that $P(y_j|x_i)$ represents the probability that y_j is received when x_i is transmitted. This is a characteristic of the channel and the receiver. Thus, a given channel (with its receiver)is specified by the **channel matrix**:

$$\begin{array}{cc} & \text{Outputs} \\ & \begin{array}{c|cccc|} & y_1 & y_2 & \cdots & y_s \\ \hline x_1 & P(y_1|x_1) & P(y_2|x_1) & \cdots & P(y_s|x_1) \\ x_2 & P(y_1|x_2) & P(y_2|x_2) & \cdots & P(y_s|x_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_r & P(y_1|x_r) & P(y_2|x_r) & \cdots & P(y_s|x_r) \end{array} \end{array}$$

$$\text{Inputs}\quad \begin{array}{c} x_1 \\ x_2 \\ \cdots \\ x_r \end{array} \begin{pmatrix} P(y_1|x_1) & P(y_2|x_1) & \cdots & P(y_s|x_1) \\ P(y_1|x_2) & P(y_2|x_2) & \cdots & P(y_s|x_2) \\ \cdots & \cdots & \cdots & \cdots \\ P(y_1|x_r) & P(y_2|x_r) & \cdots & P(y_s|x_r) \end{pmatrix}$$

We can obtain the reverse conditional probabilities $P(x_i|y_j)$ using Bayes' rule:

$$P(x_i|y_j) = \frac{P(y_j|x_i)P(x_i)}{P(y_j)} \tag{15.18a}$$

$$= \frac{P(y_j|x_i)P(x_i)}{\sum_i P(x_i, y_j)} \tag{15.18b}$$

$$= \frac{P(y_j|x_i)P(x_i)}{\sum_i P(x_i)P(y_j|x_i)} \tag{15.18c}$$

Thus, if the input symbol probabilities $P(x_i)$ and the channel matrix are known, the reverse conditional probabilities can be computed from Eqs. (15.18). The reverse conditional probability $P(x_i|y_j)$ is the probability that x_i was transmitted when y_j is received.

If the channel were noise-free, the average amount of information received would be $H(\mathrm{x})$ bits (entropy of the source) per received symbol. Note that $H(\mathrm{x})$ is the average information transmitted over the channel per symbol. Because of channel noise, we lose an average of

---

* This can be verified from the fact that for a noiseless channel all the probabilities in Eq. (15.17) are either 0 or 1. If $P(x_i|y_j) = 1$, then $\log[1/P(x_i|y_j)] = 0$ and if $P(x_i|y_j) = 0$, then $P(x_i, y_j) = P(y_j)P(x_i|y_j) = 0$. This shows that $H(\mathrm{x}|\mathrm{y}) = 0$

$H(\mathrm{x}|\mathrm{y})$ bits of information per symbol. Therefore, in this transaction the amount of information the receiver receives is, on the average, $I(\mathrm{x};\mathrm{y})$ bits per received symbol, where

$$I(\mathrm{x};\mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}) \qquad \text{bits per symbol} \tag{15.19}$$

$I(\mathrm{x};\mathrm{y})$ is called the **mutual information** of x and y. Because

$$H(\mathrm{x}) = \sum_i P(x_i) \log \frac{1}{P(x_i)} \qquad \text{bits}$$

we have

$$I(\mathrm{x};\mathrm{y}) = \sum_i P(x_i) \log \frac{1}{P(x_i)} - \sum_i \sum_j P(x_i, y_j) \log \frac{1}{P(x_i|y_j)}$$

Also because

$$\sum_j P(x_i, y_j) = P(x_i)$$

We have

$$I(\mathrm{x};\mathrm{y}) = \sum_i \sum_j P(x_i, y_j) \log \frac{1}{P(x_i)} - \sum_i \sum_j P(x_i, y_j) \log \frac{1}{P(x_i|y_j)}$$

$$= \sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i|y_j)}{P(x_i)} \tag{15.20a}$$

$$= \sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{15.20b}$$

Alternatively, using Bayes' rule in Eq. (15.20a), $I(\mathrm{x};\mathrm{y})$ may be expressed as

$$I(\mathrm{x};\mathrm{y}) = \sum_i \sum_j P(x_i, y_j) \log \frac{P(y_j|x_i)}{P(y_j)} \tag{15.20c}$$

or we may substitute Eq. (15.18c) into Eq. (15.20a):

$$I(\mathrm{x};\mathrm{y}) = \sum_i \sum_j P(x_i)P(y_j|x_i) \log \frac{P(y_j|x_i)}{\sum_i P(x_i)P(y_j|x_i)} \tag{15.20d}$$

Equation (15.20d) expresses $I(\mathrm{x};\mathrm{y})$ in terms of the input symbol probabilities and the channel matrix.

The units of $I(\mathrm{x};\mathrm{y})$ should be carefully noted. $I(\mathrm{x};\mathrm{y})$ is the average amount of information received per symbol transmitted. Hence, its units are bits per symbol. If we use binary digits at the input, then the symbol is a binary digit, and the units of $I(\mathrm{x};\mathrm{y})$ are bits per binary digit.

Because $I(\mathrm{x};\mathrm{y})$ in Eq. (15.20b) is symmetrical with respect to x and y, it follows that

$$I(\mathrm{x};\mathrm{y}) = I(\mathrm{y};\mathrm{x}) \tag{15.21a}$$

$$= H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) \tag{15.21b}$$

The quantity $H(\mathrm{y}|\mathrm{x})$ is the equivocation of y with respect to x and is the average uncertainty about the received symbol when the transmitted symbol is known. Equation (15.21b) can be rewritten as

$$H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) \tag{15.21c}$$

From Eq. (15.20d) it is clear that $I(\mathrm{x};\,\mathrm{y})$ is a function of the transmitted symbol probabilities $P(x_i)$ and the channel matrix. For a given channel, $I(\mathrm{x};\mathrm{y})$ will be maximum for some set of probabilities $P(x_i)$. This maximum value is the **channel capacity** C_s,

$$C_s = \max_{P(x_i)} I(\mathrm{x};\,\mathrm{y}) \quad \text{bits per symbol} \tag{15.22}$$

Thus, C_s represents the maximum information that can be transmitted by one symbol over the channel. These ideas will become clear from the following example of a binary symmetric channel (BSC).

EXAMPLE 15.3 Find the channel capacity of the BSC shown in Fig. 15.2.

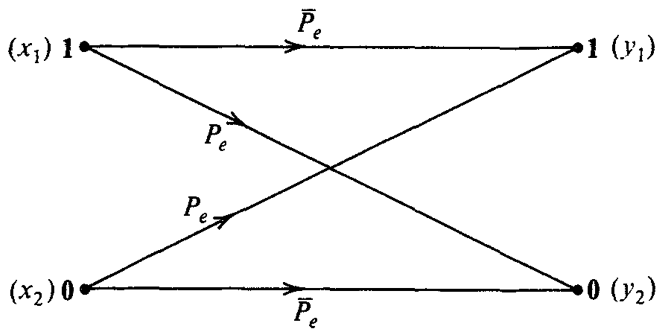


Figure 15.2 Binary symmetric channel.

Let $P(x_1) = \alpha$ and $P(x_2) = \bar{\alpha} = (1-\alpha)$. Also,

$$P(y_1|x_2) = P(y_2|x_1) = P_e$$
$$P(y_1|x_1) = P(y_2|x_2) = \bar{P}_e = 1 - P_e$$

Substitution of these probabilities into Eq. (15.20d) gives

$$\begin{aligned}
I(\mathrm{x};\,\mathrm{y}) &= \alpha\bar{P}_e \log\left(\frac{\bar{P}_e}{\alpha\bar{P}_e + \bar{\alpha}P_e}\right) + \alpha P_e \log\left(\frac{P_e}{\alpha P_e + \bar{\alpha}\bar{P}_e}\right) \\
&\quad + \bar{\alpha}P_e \log\left(\frac{P_e}{a\bar{P}_e + \bar{\alpha}P_e}\right) + \bar{\alpha}\bar{P}_e \log\left(\frac{\bar{P}_e}{\alpha P_e + \bar{\alpha}\bar{P}_e}\right) \\
&= (\alpha P_e + \bar{\alpha}\bar{P}_e)\log\left(\frac{1}{\alpha P_e + \bar{\alpha}\bar{P}_e}\right) \\
&\quad + (\alpha\bar{P}_e + \bar{\alpha}P_e)\log\left(\frac{1}{\alpha\bar{P}_e + \bar{\alpha}P_e}\right) \\
&\quad - \left(P_e \log\frac{1}{P_e} + \bar{P}_e \log\frac{1}{\bar{P}_e}\right)
\end{aligned}$$

If we define

$$\Omega(z) = z \log \frac{1}{z} + \bar{z} \log \frac{1}{\bar{z}}$$

with $\bar{z} = 1 - z$, then

$$I(\mathrm{x}; \mathrm{y}) = \Omega(\alpha P_e + \bar{\alpha}\bar{P}_e) - \Omega(P_e) \tag{15.23}$$

The function $\Omega(z)$ vs. z is shown in Fig. 15.3. It can be seen that $\Omega(z)$ is maximum at $z = \frac{1}{2}$. (Note that we are interested in the region $0 < z < 1$ only.) For a given P_e, $\Omega(P_e)$ is fixed. Hence from Eq. (15.23) it follows that $I(\mathrm{x}; \mathrm{y})$ is maximum when $\Omega(\alpha P_e + \bar{\alpha} P_e)$ is maximum. This occurs when

$$\alpha P_e + \bar{\alpha}\bar{P}_e = 0.5$$

or

$$\alpha P_e + (1 - \alpha)(1 - P_e) = 0.5$$

This equation is satisfied when

$$\alpha = 0.5 \tag{15.24}$$

For this value of α, $\Omega(\alpha P_e + \bar{\alpha}\bar{P}_e) = 1$ and

$$\begin{aligned} C_s &= \max_{P(x_i)} I(\mathrm{x}; \mathrm{y}) = 1 - \Omega(P_e) \\ &= 1 - \left[P_e \log \frac{1}{P_e} + (1 - P_e) \log \left(\frac{1}{1 - P_e}\right)\right] \end{aligned} \tag{15.25}$$

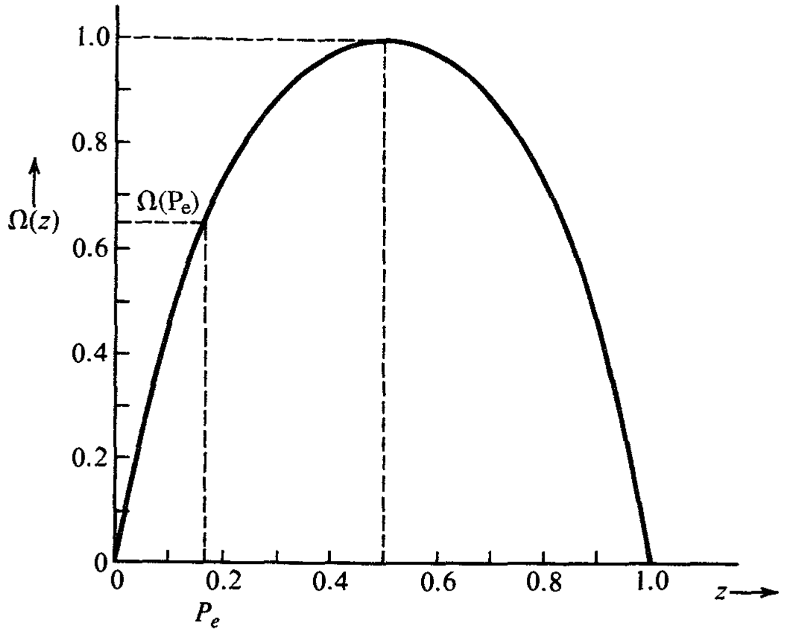


Figure 15.3 Plot of $\Omega(z)$.

Figure 15.4 shows C_s vs. P_e. From this figure it follows that the maximum value of C_s is unity. This means we can transmit at most 1 bit of information per binary digit. This is the expected result, because one binary digit can convey one of the two equiprobable messages. The information content of one of the two equiprobable messages is $\log_2 2 = 1$ bit. Second, we observe that C_s is maximum when the error probability $P_e = 0$ or $P_e = 1$. When the error probability $P_e = 0$, the channel is noiseless, and we expect C_s to be maximum. But surprisingly, C_s is also maximum when $P_e = 1$. This is easy to explain, because a channel that consistently and with certainty makes errors is as good as a noiseless channel. All we have to do is reverse the decision that is made, and we have error-free reception; that is, if **0** is received, we decide that **1** was actually sent, and vice versa. The channel capacity C_s is zero (minimum) when $P_e = \frac{1}{2}$. If the error probability is $\frac{1}{2}$, then the transmitted symbols and the received symbols are statistically independent. If we received **0**, for example, either **1** or **0** is equally likely to have been transmitted, and the information received is zero.

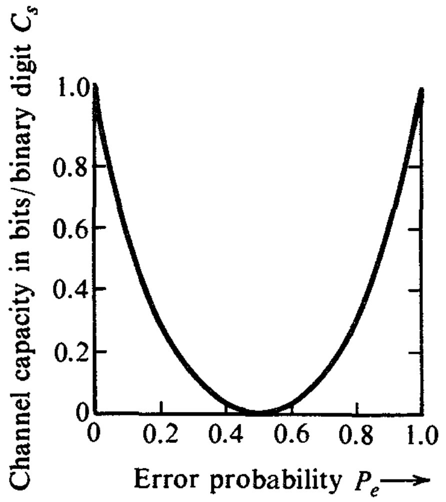


Figure 15.4 Binary symmetric channel capacity as a function of error probability P_e.

Channel Capacity per Second

The channel capacity C_s in Eq. (15.22) gives the maximum possible information transmitted when one symbol (digit) is transmitted. If K symbols are being transmitted per second, then the maximum rate of transmission of information per second is KC_s. This is the channel capacity in information units per seconds and will be denoted by C (in bits per second):

$$C = KC_s \quad \text{bit/s}$$

A Comment on Channel Capacity: Channel capacity is the property of a particular physical channel over which the information is transmitted. This is true provided the term **channel** is correctly interpreted. A channel means not only the transmission medium but it also includes the specifications of the kind of signals (binary, r-ary, etc., or orthogonal, simplex, etc.) and the kind of receiver used (the receiver determines the error probability). All these specifications are included in the channel matrix. A channel matrix completely specifies a channel. If we decide to use, for example, 4-ary digits instead of binary digits over the same physical channel, the channel matrix changes (it becomes a 4 × 4 matrix), as does the channel capacity. Similarly, a change in the receiver or the signal power or noise power will change the channel matrix and, hence, the channel capacity.

Magnitude of the Channel Capacity

The channel capacity C_s is the maximum value of $H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y})$; clearly, $C_s \leq H(\mathrm{x})$ [because $H(\mathrm{x}|\mathrm{y}) \geq 0$]. But $H(\mathrm{x})$ is the average information per input symbol. Hence, C_s is always less than (or equal to) the average information per input symbol. If we use binary symbols at the input, the maximum value of $H(\mathrm{x})$ is 1 bit, occurring when $P(x_1) = P(x_2) = \frac{1}{2}$. Hence, for a binary channel, $C_s \leq 1$ bit per binary digit. If we use r-ary symbols, the maximum value of $H_r(\mathrm{x})$ is 1 r-ary unit. Hence, $C_s \leq 1$ r-ary unit per symbol.

Verification of Error-Free Communication over a BSC

We have shown that over a noisy channel, C_s bits of information can be transmitted per symbol. If we consider a binary channel, this means that for each binary digit (symbol) transmitted, the received information is C_s bits ($C_s \leq 1$). Thus, to transmit 1 bit of information, we need to transmit at least $1/C_s$ binary digits. This gives a code efficiency C_s and redundancy $1 - C_s$. When the transmission of information is implied, it means error-free transmission, because $I(\mathrm{x}; \mathrm{y})$ was defined as the transmitted information minus the loss of information caused by channel noise.

The problem with this derivation is that it is based on a certain speculative definition of information [Eq. (15.1)]. And based on this definition, we defined the information lost during the transmission over the channel. We really have no direct proof that the information lost over the channel will oblige us in this way. Hence, the only way to ensure that this whole speculative structure is sound is to verify it. If we can show that C_s bits of error-free information can be transmitted per symbol over a channel, the verification will be complete. A general case will be discussed later. Here we shall verify the results for a BSC.

Let us consider a binary source emitting messages at a rate of α digits per second. We accumulate these information digits over T seconds to give a total of αT digits. Because αT digits form $2^{\alpha T}$ possible combinations, our problem is now to transmit one of these $2^{\alpha T}$ supermessages every T seconds. These supermessages are transmitted by a code of word length βT digits, with $\beta > \alpha$ to ensure redundancy. Because βT digits can form $2^{\beta T}$ distinct patterns (vertices of a βT-dimensional hypercube), and we have only $2^{\alpha T}$ messages, we are utilizing only a $2^{-(\beta-\alpha)T}$ fraction of the vertices. The remaining vertices are deliberately unused in order to combat noise. If we let $T \to \infty$, the fraction of vertices used approaches 0. Because there are βT digits in each transmitted sequence, the number of digits received in error will be exactly $\beta T P_e$ when $T \to \infty$. We now construct Hamming spheres of radius $\beta T P_e$ each around the $2^{\alpha T}$ vertices used for the messages. When any message is transmitted, the received message will be on the Hamming sphere surrounding the vertex corresponding to that message. We use the following decision rule: If a received sequence falls inside or on a sphere surrounding message m_i, then the decision is "m_i is transmitted." If $T \to \infty$, the decision will be without error if all the $2^{\alpha T}$ spheres are nonoverlapping.

Of all the possible sequences of βT digits, the number of sequences that differ from a given sequence by exactly j digits is $\binom{\beta T}{j}$ (see Example 10.6). Hence, K, the total number of sequences that differ from a given sequence by less than or equal to $\beta T P_e$ digits, is

$$K = \sum_{j=0}^{\beta T P_e} \binom{\beta T}{j} \tag{15.26}$$

Here we use an inequality often used in information theory:[4,7]

$$\sum_{j=0}^{\beta T P_e} \binom{\beta T}{j} \le 2^{\beta T \Omega(P_e)} \qquad P_e < 0.5$$

Hence,

$$K \le 2^{\beta T \Omega(P_e)} \tag{15.27}$$

with

$$\Omega(P_e) = P_e \log \frac{1}{P_e} + (1 - P_e) \log \frac{1}{1 - P_e}$$

From the $2^{\beta T}$ possible vertices we choose $2^{\alpha T}$ vertices to be assigned to the supermessages. How shall we select these vertices? From the decision procedure it is clear that if we assign a particular vertex to a supermessage, then none of the other vertices lying within a sphere of radius $\beta T P_e$ can be assigned to another supermessage. Thus, when we choose a vertex for m_1, the corresponding K vertices [Eq. (15.26)] become ineligible for consideration. From the remaining $2^{\beta T} - K$ vertices we choose another vertex for m_2. We proceed this way until all the $2^{\beta T}$ vertices are exhausted. This is a rather tedious procedure. Let us see what happens if we choose the required $2^{\alpha T}$ vertices randomly from the $2^{\beta T}$ vertices. In this procedure there is a danger that we may select more than one vertex lying within a distance $\beta T P_e$. If, however, α/β is sufficiently small, the probability of making such a choice is extremely small as $T \to \infty$. The probability of choosing any particular vertex s_1 as one of the $2^{\alpha T}$ vertices from $2^{\beta T}$ vertices is $2^{\alpha T}/2^{\beta T} = 2^{-(\beta-\alpha)T}$.

Remembering that K vertices lie within a distance of $\beta T P_e$ digits from s_1, the probability that we may also choose another vertex s_2 that is within the distance $\beta T P_e$ from s_1 is

$$P = K 2^{-(\beta-\alpha)T}$$

From Eq. (15.27) it follows that

$$P \le 2^{-\{\beta[1-\Omega(P_e)]-\alpha\}T}$$

Hence, as $T \to \infty$, $P \to 0$ if

$$\beta[1 - \Omega(P_e)] > \alpha$$

that is, if

$$\frac{\alpha}{\beta} < 1 - \Omega(P_e) \tag{15.28a}$$

But $1 - \Omega(P_e)$ is C_s, the channel capacity of a BSC [Eq. (15.25)]. Therefore,

$$\frac{\alpha}{\beta} < C_s \tag{15.28b}$$

Hence, the probability of choosing two sequences randomly within a distance $\beta T P_e$ approaches 0 as $T \to \infty$ provided $\alpha/\beta < C_s$, and we have error-free communication. We can choose $\alpha/\beta = C_s - \epsilon$, where ϵ is arbitrarily small.

15.5 CHANNEL CAPACITY OF A CONTINUOUS CHANNEL*

For a discrete random variable x taking on values $x_1, x_2, \ldots, x_n$ with probabilities $P(x_1)$, $P(x_2), \ldots, P(x_n)$, the entropy $H(\mathrm{x})$ was defined as

$$H(\mathrm{x}) = \sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{15.29}$$

For analog data, we have to deal with continuous random variables. Therefore, we must extend the definition of entropy to continuous random variables. One is tempted to state that $H(\mathrm{x})$ for continuous random variables is obtained by using the integral instead of discrete summation in Eq. (15.29)†:

$$H(\mathrm{x}) = \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx \tag{15.30}$$

We shall see that Eq. (15.30) is indeed the meaningful definition of entropy for a continuous random variable. We cannot accept this definition, however, unless we show that it has the meaningful interpretation as uncertainty. A random variable x takes a value in the range $(n\Delta x,\ (n+1)\Delta x)$ with probability $p(n\Delta x)\Delta x$ in the limit as $\Delta x \to 0$. The error in the approximation will vanish in the limit as $\Delta x \to 0$. Hence $H(\mathrm{x})$, the entropy of a continuous random variable x, is given by

$$\begin{aligned}
H(\mathrm{x}) &= \lim_{\Delta x \to 0} \sum_n p(n\Delta x)\Delta x \log \frac{1}{p(n\Delta x)\Delta x} \\
&= \lim_{\Delta x \to 0} \left[\sum_n p(n\Delta x)\Delta x \log \frac{1}{p(n\Delta x)} - \sum_n p(n\Delta x)\Delta x \log \Delta x \right] \\
&= \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx - \lim_{\Delta x \to 0} \log \Delta x \int_{-\infty}^{\infty} p(x)\, dx \\
&= \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx - \lim_{\Delta x \to 0} \log \Delta x
\end{aligned} \tag{15.31}$$

In the limit as $\Delta x \to 0$, $\log \Delta x \to -\infty$. It therefore appears that the entropy of a continuous random variable is infinite. This is quite true. The magnitude of uncertainty associated with a continuous random variable is infinite. This fact is also apparent intuitively. A continuous random variable assumes a nonenumerably infinite number of values, and, hence, the uncertainty is on the order of infinity. Does this mean that there is no meaningful definition of entropy for a continuous random variable? On the contrary, we shall see that the first term in Eq. (15.31) serves as a meaningful measure of the entropy (average information) of a continuous random variable x. This may be argued as follows. We can consider $\int p(x) \log [1/p(x)]\, dx$ as a relative entropy with $-\log \Delta x$ serving as a datum, or reference. The information transmitted over a channel is actually the difference between the two terms $H(\mathrm{x})$ and $H(\mathrm{x}|\mathrm{y})$. Obviously, if we have a common datum for both $H(\mathrm{x})$ and $H(\mathrm{x}|\mathrm{y})$, the difference $H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y})$ will be the same as the difference between their relative entropies. We are therefore justified in considering the first term in Eq. (15.31) as the **differential** entropy of x. We must, however, always remember that this is a relative entropy and not the absolute entropy. Failure to realize this subtle point generates many apparent fallacies, one of which is given in Example 15.4.

---

* The channel is assumed to be memoriless.

† Throughout this discussion, the PDF $p_{\mathrm{x}}(x)$ will be abbreviated as $p(x)$, because it causes no ambiguity and improves the clarity of equations.

Based on this argument, we define $H(\mathrm{x})$, the differential entropy of a continuous random variable x, as

$$H(\mathrm{x}) = \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)}\, dx \quad \text{bits} \tag{15.32a}$$

$$= -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx \quad \text{bits} \tag{15.32b}$$

Although $H(\mathrm{x})$ is the differential (relative) entropy of x, we shall call it the entropy of random variable x.

EXAMPLE 15.4 A signal amplitude x is a random variable uniformly distributed in the range $(-1, 1)$. This signal is passed through an amplifier of gain 2. The output y is also a random variable, uniformly distributed in the range $(-2, 2)$. Determine the (differential) entropies $H(\mathrm{x})$ and $H(\mathrm{y})$.

We have

$$P(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(y) = \begin{cases} \frac{1}{4} & |y| < 2 \\ 0 & \text{otherwise} \end{cases}$$

Hence,

$$H(\mathrm{x}) = \int_{-1}^{1} \frac{1}{2} \log 2\, dx = 1 \text{ bit}$$

$$H(\mathrm{y}) = \int_{-2}^{2} \frac{1}{4} \log 4\, dx = 2 \text{ bits}$$

The entropy of the random variable y is twice that of x. This result may come as a surprise, since a knowledge of x uniquely determines y, and vice versa, because y = 2x. Hence, the average uncertainty of x and y should be identical. Amplification itself can neither add nor subtract information. Why, then, is $H(\mathrm{y})$ twice as large as $H(\mathrm{x})$? This becomes clear when we remember that $H(\mathrm{x})$ and $H(\mathrm{y})$ are differential (relative) entropies, and they will be equal if and only if their datum (or reference) entropies are equal. The reference entropy R_1 for x is $-\log \Delta x$, and the reference entropy R_2 for y is $-\log \Delta y$ (in the limit as $\Delta x,\ \Delta y \to 0$),

$$R_1 = \lim_{\Delta x \to 0} -\log \Delta x$$

$$R_2 = \lim_{\Delta y \to 0} -\log \Delta y$$

and

$$R_1 - R_2 = \lim_{\Delta x, \Delta y \to 0} \log \left(\frac{\Delta y}{\Delta x} \right)$$

$$= \log \left(\frac{dy}{dx} \right)$$

$$= \log 2 = 1 \text{ bit}$$

Thus, R_1, the reference entropy of x, is higher than the reference entropy R_2 for y. Hence, if x and y have equal absolute entropies, their differential (relative) entropies must differ by 1 bit.

Maximum Entropy for a Given Mean Square Value of x

For discrete random variables, we observed that the entropy is maximum when all the outcomes (messages) were equally likely (uniform probability distribution). For continuous random variables, there also exists a PDF $p(x)$ that maximizes $H(\mathrm{x})$ in Eqs. (15.32). In the case of a continuous distribution, however, we may have additional constraints on x. Either the maximum value of x or the mean square value of x may be given. We shall find here the PDF $p(x)$ that will yield maximum entropy when $\overline{\mathrm{x}^2}$ is given to be a constant σ^2. The problem, then, is to maximize $H(\mathrm{x})$:

$$H(\mathrm{x}) = \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)}\, dx \tag{15.33}$$

with the constraints

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{15.34a}$$

$$\int_{-\infty}^{\infty} x^2 p(x)\, dx = \sigma^2 \tag{15.34b}$$

To solve this problem, we use a theorem from the calculus of variation. Given the integral I,

$$I \int_a^b F(x, p)\, dx \tag{15.35}$$

subject to the following constraints:

$$\int_a^b \varphi_1(x, p)\, dx = \lambda_1$$

$$\int_a^b \varphi_2(x, p)\, dx = \lambda_2$$

$$\vdots$$

$$\int_a^b \varphi_k(x, p)\, dx = \lambda_k \tag{15.36}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are given constants. The result from the calculus of variation states that the form of $p(x)$ that maximizes I in Eq. (15.35) with the constraints in Eq. (15.36) is found from the solution of the equation

$$\frac{\partial F}{\partial p} + \alpha_1 \frac{\partial \varphi_1}{\partial p} + \alpha_2 \frac{\partial \varphi_2}{\partial p} + \cdots + \alpha_k \frac{\partial \varphi_k}{\partial p} = 0 \tag{15.37}$$

The quantities $\alpha_1, \alpha_2, \ldots, \alpha_k$ are adjustable constants, called **undetermined multipliers**, which can be found by substituting the solution of $p(x)$ [obtained from Eq. (15.37)] in Eq. (15.36). In the present case,

$$F(p, x) = p \log \frac{1}{p}$$

$$\varphi_1(x, p) = p$$

$$\varphi_2(x, p) = x^2 p$$

Hence, the solution for p is given by

$$\frac{\partial}{\partial p}\left(p \log \frac{1}{p}\right) + \alpha_1 + \alpha_2 \frac{\partial}{\partial p} x^2 p = 0$$

or

$$-(1 + \log p) + \alpha_1 + \alpha_2 x^2 = 0$$

Solving for p, we have

$$p = e^{(\alpha_1 - 1)} e^{\alpha_2 x^2} \tag{15.38}$$

Substituting Eq. (15.38) into Eq. (15.34a), we have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} e^{\alpha_1 - 1} e^{\alpha_2 x^2} dx \\ &= 2e^{\alpha_1 - 1} \int_0^{\infty} e^{\alpha_2 x^2} dx \\ &= 2e^{\alpha_1 - 1} \left(\frac{1}{2}\sqrt{\frac{\pi}{-\alpha_2}}\right) \end{aligned}$$

provided α_2 is negative, or

$$e^{\alpha_1 - 1} = \sqrt{\frac{-\alpha_2}{\pi}} \tag{15.39}$$

Next we substitute Eqs. (15.38) and (15.39) into Eq. (15.34b):

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} x^2 \sqrt{\frac{-\alpha_2}{\pi}} e^{\alpha_2 x^2} dx \\ &= 2\sqrt{\frac{-\alpha_2}{\pi}} \int_0^{\infty} x^2 e^{\alpha_2 x^2} dx \\ &= -\frac{1}{2\alpha_2} \end{aligned}$$

or

$$\alpha_2 = -\frac{1}{2\sigma^2} \tag{15.40a}$$

and

$$e^{\alpha_1 - 1} = \sqrt{\frac{1}{2\pi\sigma^2}} \tag{15.40b}$$

Substituting Eqs. (15.40) into Eq. (15.38), we have

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \tag{15.41}$$

We therefore conclude that for a given mean square value, the maximum entropy (or maximum uncertainty) is obtained when the distribution of x is gaussian. This maximum entropy, or uncertainty, is given by

$$H(\mathrm{x}) = \int_{-\infty}^{\infty} p(x) \log_2 \frac{1}{p(x)}\, dx$$

Note that

$$\begin{aligned}
\log \frac{1}{p(x)} &= \log\left(\sqrt{2\pi\sigma^2}\, e^{x^2/2\sigma^2}\right) \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{x^2}{2\sigma^2}\log e
\end{aligned}$$

Hence,

$$\begin{aligned}
H(x) &= \int_{-\infty}^{\infty} p(x)\left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{x^2}{2\sigma^2}\log e\right] dx \\
&= \frac{1}{2}\log(2\pi\sigma^2)\int_{-\infty}^{\infty} p(x)\, dx + \frac{\log e}{2\sigma^2}\int_{-\infty}^{\infty} x^2 p(x)\, dx \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\log e}{2\sigma^2}\sigma^2 \\
&= \frac{1}{2}\log(2\pi e\sigma^2) && \text{(15.42a)} \\
&= \frac{1}{2}\log(17.1\sigma^2) && \text{(15.42b)}
\end{aligned}$$

To reiterate, for a given mean square value $\overline{\mathrm{x}^2}$, the entropy is maximum for a gaussian distribution, and the corresponding entropy is $\frac{1}{2}\log(2\pi e\sigma^2)$.

The reader can similarly show that if x is constrained to some peak value M $(-M < \mathrm{x} < M)$, then the entropy is maximum when x is uniformly distributed:

$$p(x) = \begin{cases} \frac{1}{2}M & -M < x < M \\ 0 & \text{otherwise} \end{cases}$$

Entropy of a Band-Limited White Gaussian Noise

Consider a band-limited white gaussian noise $\mathrm{n}(t)$ with PSD $\mathcal{N}/2$. Because

$$R_{\mathrm{n}}(\tau) = \mathcal{N}B \operatorname{sinc}(2\pi B\tau)$$

we know that $\operatorname{sinc}(2\pi B\tau)$ is zero at $\tau = \pm k/2B$ (k integer). Therefore,

$$R_{\mathrm{n}}\left(\frac{k}{2B}\right) = 0 \qquad k = \pm 1, \pm 2, \pm 3, \ldots$$

Hence,

$$R_{\mathrm{n}}\left(\frac{k}{2B}\right) = \overline{\mathrm{n}(t)\mathrm{n}\left(t + \frac{k}{2B}\right)} = 0 \qquad k = \pm 1, \pm 2, \ldots$$

Because $\mathrm{n}(t)$ and $\mathrm{n}(t + k/2B)$ ($k = \pm 1, \pm 2, \ldots$) are Nyquist samples of $\mathrm{n}(t)$, it follows that Nyquist samples of $\mathrm{n}(t)$ are all uncorrelated. Because $\mathrm{n}(t)$ is gaussian, uncorrelatedness implies independence. Hence, all Nyquist samples of $\mathrm{n}(t)$ are independent. Note that

$$\overline{\mathrm{n}^2} = R_{\mathrm{n}}(0) = \mathcal{N}B$$

Hence, the variance of each Nyquist sample is $\mathcal{N}B$. From Eq. (15.42a) it follows that the entropy $H(\mathrm{n})$ of each Nyquist sample of $\mathrm{n}(t)$ is

$$H(\mathrm{n}) = \frac{1}{2}\log(2\pi e\mathcal{N}B) \quad \text{bits per sample} \tag{15.43a}$$

Because $\mathrm{n}(t)$ is completely specified by $2B$ Nyquist samples per second, the entropy per second of $\mathrm{n}(t)$ is the entropy of $2B$ Nyquist samples. Because all the samples are independent, knowledge of one sample gives no information about any other sample. Hence, the entropy of $2B$ Nyquist samples is the sum of the entropies of the $2B$ samples, and

$$H'(\mathrm{n}) = B\log(2\pi e\mathcal{N}B) \quad \text{bit/s} \tag{15.43b}$$

where $H'(\mathrm{n})$ is the entropy per second of $\mathrm{n}(t)$.

From the results derived thus far, we can draw one significant conclusion. Among all signals band-limited to B Hz and constrained to have a certain mean square value σ^2, the white gaussian band-limited signal has the largest entropy per second. The reason for this lies in the fact that for a given mean square value, gaussian samples have the largest entropy; moreover, all the $2B$ samples of a gaussian band-limited process are independent. Hence, the entropy per second is the sum of the entropies of all the $2B$ samples. In processes that are not white, the Nyquist samples are correlated, and, hence, the entropy per second is less than the sum of the entropies of the $2B$ samples. If the signal is not gaussian, then its samples are not gaussian, and, hence, the entropy per sample is also less than the maximum possible entropy for a given mean square value. To reiterate, for a class of band-limited signals constrained to a certain mean square value, the white gaussian signal has the largest entropy per second, or the largest amount of uncertainty. This is also the reason why white gaussian noise is the worst possible noise in terms of interference with signal transmission.

Mutual Information $I(\mathrm{x}; \mathrm{y})$

The ultimate test of any concept is its usefulness. We shall now show that the relative entropy defined in Eqs. (15.32) does lead to meaningful results when we consider $I(\mathrm{x}; \mathrm{y})$, the mutual information of continuous random variables x and y. We wish to transmit a random variable x over a channel. Each value of x in a given continuous range is now a message that may be transmitted, for example, as a pulse of height x. The message recovered by the receiver will be a continuous random variable y. If the channel were noise-free, the received value y would uniquely determine the transmitted value x. But channel noise introduces a certain uncertainty about the true value of x. Consider the event that at the transmitter, a value of x in the interval $(x, x + \Delta x)$ has been transmitted $(\Delta x \to 0)$. The probability of this event is $p(x)\Delta x$ in the limit $\Delta x \to 0$. Hence, the amount of information transmitted is $\log[1/p(x)\Delta x]$. Let the value of y at the receiver be y and let $p(x|y)$ be the conditional probability density of x when y = y. Then $p(x|y)\Delta x$ is the probability that x will lie in the interval $(x, x + \Delta x)$ when y = y (provided $\Delta x \to 0$). Obviously, there is an uncertainty about the event that x lies in the interval $(x, x + \Delta x)$. This uncertainty, $\log[1/p(x|y)\Delta x]$, arises because of channel noise and therefore represents a loss of information. Because $\log[1/p(x)\Delta x]$ is the information transmitted and $\log[1/p(x|y)\Delta x]$ is the information lost over the channel, the net information received is $I(x; y)$ given by

$$I(x; y) = \log \frac{p(x|y)}{p(x)} \tag{15.44}$$

Note that this relation is true in the limit $\Delta x \to 0$. Therefore, $I(x; y)$, represents the information transmitted over a channel when we receive y (y = y) when x is transmitted (x = x). We are interested in finding the average information transmitted over a channel when some x is transmitted and a certain y is received. We must therefore average $I(x; y)$ over all values of x and y. The average information transmitted will be denoted by $I(\mathrm{x}; \mathrm{y})$, where

$$I(\mathrm{x}; \mathrm{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) I(x; y)\, dx\, dy \tag{15.45a}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \log \frac{p(x|y)}{p(x)}\, dx\, dy \tag{15.45b}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \log \frac{1}{p(x)}\, dx\, dy \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \log p(x|y)\, dx\, dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y|x) \log \frac{1}{p(x)}\, dx\, dy \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \log p(x|y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx \int_{-\infty}^{\infty} p(y|x)\, dy \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x, y) \log p(x|y)\, dx\, dy
\end{aligned}$$

Note that

$$\int_{-\infty}^{\infty} p(y|x)\, dy = 1 \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)}\, dx = H(\mathrm{x})$$

Hence,

$$I(\mathrm{x};\mathrm{y}) = H(\mathrm{x}) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\log p(x|y)\,dx\,dy \tag{15.46a}$$

$$= H(\mathrm{x}) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\log \frac{1}{p(x|y)}\,dx\,dy \tag{15.46b}$$

The integral on the right-hand side is the average over x and y of $\log[1/p(x|y)]$. But $\log[1/p(x|y)]$ represents the uncertainty about x when y is received. This, as we have seen, is the information lost over the channel. The average of $\log[1/p(x|y)]$ is the average loss of information when some x is transmitted and some y is received. This, by definition, is $H(\mathrm{x}|\mathrm{y})$, the equivocation of x with respect to y,

$$H(\mathrm{x}|\mathrm{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\log \frac{1}{p(x|y)}\,dx\,dy \tag{15.47}$$

Hence,

$$I(\mathrm{x};\mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}) \tag{15.48}$$

Thus, when some value of x is transmitted and some value of y is received, the average information transmitted over the channel is $I(\mathrm{x};\mathrm{y})$, given by Eq. (15.48). We can define the channel capacity C_s as the maximum amount of information that can be transmitted, on the average, per sample or per value transmitted:

$$C_s = \max I(\mathrm{x};\mathrm{y}) \tag{15.49}$$

For a given channel, $I(\mathrm{x};\mathrm{y})$ is a function of the input probability density $p(x)$ alone. This can be shown as follows:

$$p(x,y) = p(x)p(y|x) \tag{15.50}$$

$$\begin{aligned}\frac{p(x|y)}{p(x)} &= \frac{p(y|x)}{p(y)}\\ &= \frac{p(y|x)}{\int_{-\infty}^{\infty} p(x,y)dx}\\ &= \frac{p(y|x)}{\int_{-\infty}^{\infty} p(x)p(y|x)dx}\end{aligned} \tag{15.51}$$

Substituting Eqs. (15.50) and (15.51) into Eq. (15.45b), we obtain

$$I(\mathrm{x};\mathrm{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y|x)\log\left(\frac{p(y|x)}{\int_{-\infty}^{\infty} p(x)p(y|x)\,dx}\right)dx\,dy \tag{15.52}$$

The conditional probability density $p(y|x)$ is characteristic of a given channel. Hence, for a given channel, $I(\mathrm{x};\mathrm{y})$ is a function of the input probability density $p(x)$ alone. Thus,

$$C_s = \max_{p(x)} I(\mathrm{x};\mathrm{y})$$

If the channel allows the transmission of K values per second, then C, the channel capacity per second, is given by

$$C = KC_s \text{ bit/s} \tag{15.53}$$

Just as in the case of discrete variables, $I(\mathrm{x}; \mathrm{y})$ is symmetrical with respect to x and y for continuous random variables. This can be seen by rewriting Eq. (15.45b) as

$$I(\mathrm{x}; \mathrm{y}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}\, dx\, dy \tag{15.54}$$

This equation shows that $I(\mathrm{x}; \mathrm{y})$ is symmetrical with respect to x and y. Hence,

$$I(\mathrm{x}; \mathrm{y}) = I(\mathrm{y}; \mathrm{x})$$

From Eq. (15.48) it now follows that

$$I(\mathrm{x}; \mathrm{y}) = H(\mathrm{x}) - H(\mathrm{x}|\mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) \tag{15.55}$$

Capacity of a Band-Limited AWGN Channel

The channel capacity C is, by definition, the maximum rate of information transmission over a channel. The mutual information $I(\mathrm{x}; \mathrm{y})$ is given by Eq. (15.55):

$$I(\mathrm{x}; \mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x}) \tag{15.56}$$

The channel capacity C is the maximum value of the mutual information $I(\mathrm{x}; \mathrm{y})$ per second. Let us first find the maximum value of $I(\mathrm{x}; \mathrm{y})$ per sample. We shall find here the capacity of a channel band-limited to B Hz and disturbed by a white gaussian noise of PSD $\mathcal{N}/2$. In addition, we shall constrain the signal power (or its mean square value) to S. The disturbance is assumed to be additive; that is, the received signal $\mathrm{y}(t)$ is given by

$$\mathrm{y}(t) = \mathrm{x}(t) + \mathrm{n}(t) \tag{15.57}$$

Because the channel is band-limited, both the signal $\mathrm{x}(t)$ and the noise $\mathrm{n}(t)$ are band-limited to B Hz. Obviously, $\mathrm{y}(t)$ is also band-limited to B Hz. All these signals can therefore be completely specified by samples taken at the uniform rate of $2B$ samples per second. Let us find the maximum information that can be transmitted per sample. Let x, n, and y represent samples of $\mathrm{x}(t)$, $\mathrm{n}(t)$, and $\mathrm{y}(t)$, respectively. The information $I(\mathrm{x}; \mathrm{y})$ transmitted per sample is given by Eq. (15.56):

$$I(\mathrm{x}; \mathrm{y}) = H(\mathrm{y}) - H(\mathrm{y}|\mathrm{x})$$

We shall now find $H(\mathrm{y}|\mathrm{x})$. By definition [Eq. (15.47)],

$$\begin{aligned} H(\mathrm{y}|\mathrm{x}) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log \frac{1}{p(y|x)}\, dx\, dy \\ &= \int_{-\infty}^{\infty} p(x)\, dx \int_{-\infty}^{\infty} p(y|x) \log \frac{1}{p(y|x)}\, dy \end{aligned}$$

Because

$$y = x + n$$

for a given x, y is equal to n plus a constant (x). Hence, the distribution of y when x has a given value is identical to that of n except for a translation by x. If $p_{\mathrm{n}}(\cdot)$ represents the PDF of noise sample n, then

$$p(y|x) = p_{\mathrm{n}}(y - x) \tag{15.58}$$

$$\int_{-\infty}^{\infty} p(y|x) \log \frac{1}{p(y|x)}\, dy = \int_{-\infty}^{\infty} p_{\mathrm{n}}(y - x) \log \frac{1}{p_{\mathrm{n}}(y - x)}\, dy$$

Letting $y - x = z$, we have

$$\int_{-\infty}^{\infty} p(y|x) \log \frac{1}{p(y|x)}\, dy = \int_{-\infty}^{\infty} p_{\mathrm{n}}(z) \log \frac{1}{p_{\mathrm{n}}(z)}\, dz$$

The right-hand side is the entropy $H(\mathrm{n})$ of the noise sample n. Hence,

$$\begin{aligned} H(\mathrm{y}|\mathrm{x}) &= H(\mathrm{n}) \int_{-\infty}^{\infty} p(x)\, dx \\ &= H(\mathrm{n}) \end{aligned} \tag{15.59}$$

In deriving Eq. (15.59), we made no assumptions about the noise. Hence, Eq. (15.59) is very general and applies to all types of noise. The only condition is that the noise disturb the channel in an additive fashion. Thus,

$$I(\mathrm{x}; \mathrm{y}) = H(\mathrm{y}) - H(\mathrm{n}) \quad \text{bits per sample} \tag{15.60}$$

We have assumed that the mean square value of the signal $\mathrm{x}(t)$ is constrained to have a value S, and the mean square value of the noise is N. We shall also assume that the signal $\mathrm{x}(t)$ and the noise $\mathrm{n}(t)$ are independent. In such a case, the mean square value of y will be the sum of the mean square values of x and n. Hence,

$$\overline{\mathrm{y}^2} = S + N$$

For a given noise [given $H(\mathrm{n})$], $I(\mathrm{x}; \mathrm{y})$ is maximum when $H(\mathrm{y})$ is maximum. We have seen that for a given mean square value of y ($\overline{\mathrm{y}^2} = S + N$), $H(\mathrm{y})$ will be maximum if y is gaussian, and the maximum entropy $H_{\max}(\mathrm{y})$ is then given by

$$H_{\max}(\mathrm{y}) = \frac{1}{2} \log\,[2\pi e(S + N)] \tag{15.61}$$

Because

$$\mathrm{y} = \mathrm{x} + \mathrm{n}$$

and n is gaussian, y will be gaussian only if x is gaussian. As the mean square value of x is S, this implies that

$$p(x) = \frac{1}{\sqrt{2\pi S}} e^{-x^2/2S}$$

and

$$\begin{aligned} I_{\max}(\mathrm{x}; \mathrm{y}) &= H_{\max}(\mathrm{y}) - H(\mathrm{n}) \\ &= \frac{1}{2} \log\,[2\pi e(S + N)] - H(\mathrm{n}) \end{aligned}$$

For a white gaussian noise with mean square value N,

$$H(\mathrm{n}) = \frac{1}{2}\log 2\pi e N \qquad N = \mathcal{N}B$$

and

$$C_s = I_{\max}(\mathrm{x};\, \mathrm{y}) = \frac{1}{2}\log\left(\frac{S+N}{N}\right) \tag{15.62a}$$

$$= \frac{1}{2}\log\left(1+\frac{S}{N}\right) \tag{15.62b}$$

The channel capacity per second will be the maximum information that can be transmitted per second. Equations (15.62) represent the maximum information transmitted per sample. If all the samples are statistically independent, the total information transmitted per second will be $2B$ times C_s. If the samples are not independent, then the total information will be less than $2BC_s$. Because the channel capacity C represents the maximum possible information transmitted per second,

$$C = 2B\left[\frac{1}{2}\log\left(1+\frac{S}{N}\right)\right]$$

$$= B\log\left(1+\frac{S}{N}\right) \quad \text{bit/s} \tag{15.63}$$

The samples of a band-limited gaussian signal are independent if and only if the signal PSD is uniform over the band (see Example 11.2 and Prob. 11.2-3). Obviously, to transmit information at the maximum rate $\lfloor$Eq. (15.63)$\rfloor$, the PSD of signal $\mathrm{y}(t)$ must be uniform. The PSD of y is given by

$$S_{\mathrm{y}}(\omega) = S_{\mathrm{x}}(\omega) + S_{\mathrm{n}}(\omega)$$

Because $S_{\mathrm{n}}(\omega) = \mathcal{N}/2$, the PSD of $\mathrm{x}(t)$ must also be uniform. Thus, the maximum rate of transmission (C bit/s) is attained when $\mathrm{x}(t)$ is also a white gaussian signal.

To recapitulate, when the channel noise is additive, white, gaussian with mean square value N ($N = \mathcal{N}B$), the channel capacity C of a band-limited channel under the constraint of a given signal power S is given by

$$C = B\log\left(1+\frac{S}{N}\right) \qquad \text{bit/s}$$

where B is the channel bandwidth in hertz. The maximum rate of transmission (C bit/s) can be realized only if the input signal is a white gaussian signal.

Capacity of a Channel of Infinite Bandwidth: Superficially, Eq. (15.63) seems to indicate that the channel capacity goes to ∞ as the channel's bandwidth B goes to ∞. This, however, is not true. For white noise, the noise power $N = \mathcal{N}B$. Hence, as B increases, N also increases. It can be shown that in the limit as $B \to \infty$, C approaches a limit:

$$C = B\log\left(1+\frac{S}{N}\right)$$

$$= B\log\left(1+\frac{S}{\mathcal{N}B}\right)$$

$$\lim_{B\to\infty} C = \lim_{B\to\infty} B \log\left(1+\frac{S}{\mathcal{N}B}\right)$$
$$= \lim_{B\to\infty} \frac{S}{\mathcal{N}}\left[\frac{\mathcal{N}B}{S}\log\left(1+\frac{S}{\mathcal{N}B}\right)\right]$$

This limit can be found by noting that

$$\lim_{x\to\infty} x\log_2\left(1+\frac{1}{x}\right) = \log_2 e = 1.44$$

Hence,

$$\lim_{B\to\infty} C = 1.44\frac{S}{\mathcal{N}} \quad \text{bit/s} \tag{15.64}$$

Thus, for a white gaussian channel noise, the channel capacity C approaches a limit of $1.44S/\mathcal{N}$ as $B \to \infty$. The variation of C with B is shown in Fig. 15.5. It is evident that the capacity can be made infinite only by increasing the signal power S to infinity. For finite signal and noise powers, the channel capacity always remains finite.

Verification of Error-Free Communication over a Continuous Channel

Using the concepts of information theory, we have shown that it is possible to transmit error-free information at a rate of $B\log_2(1+S/N)$ bit/s over a channel band-limited to B Hz. The signal power is S, and the channel noise is white gaussian with power N. This theorem can be verified in a way similar to that used for the verification of the channel capacity of a discrete case. This verification using signal space is so general that it is in reality an alternate proof of the capacity theorem.

Let us consider M-ary communication with M equiprobable messages $m_1, m_2, \ldots, m_M$ transmitted by signals $s_1(t), s_2(t), \ldots, s_M(t)$. All signals are time-limited with duration T and have an essential bandwidth B Hz. Their powers are less than or equal to S. The channel is band-limited to B, and the channel noise is white gaussian with power N.

All the signals and noise waveforms have $2BT+1$ dimensions. In the limit we shall let $T\to\infty$. Hence $2BT \gg 1$, and the number of dimensions will be taken as $2BT$ in our future discussion. Because the noise power is N, the energy of the noise waveform of T-second duration is NT. The maximum signal energy is ST. Because signals and noise are independent, the maximum received energy is $(S+N)T$. Hence, all the received signals will lie in a $2BT$-dimensional hypersphere of radius $\sqrt{(S+N)T}$ (Fig. 15.6a). A typical received

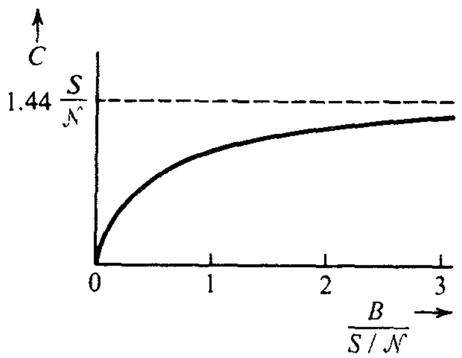


Figure 15.5 Channel capacity vs. bandwidth for a channel with white gaussian noise and fixed power.

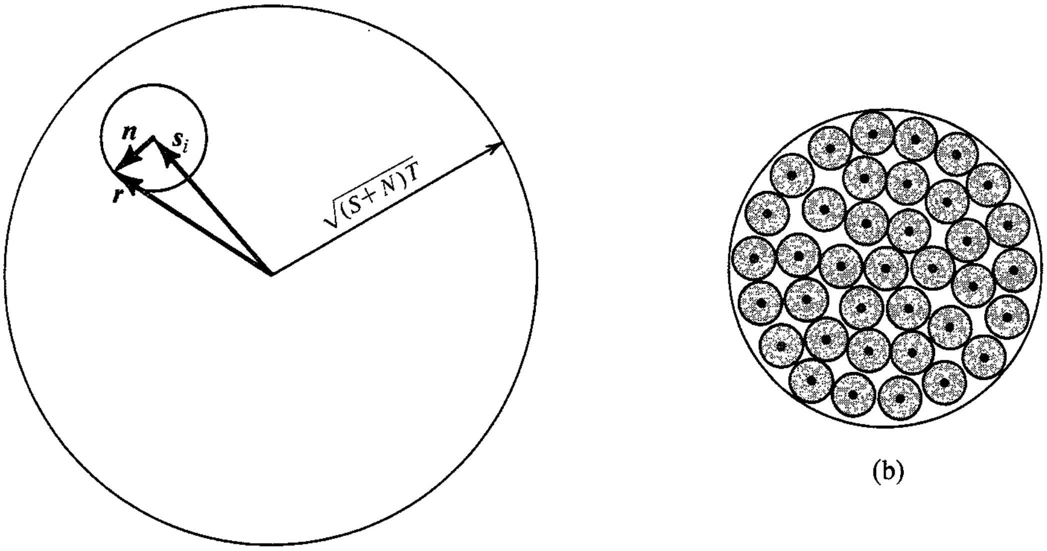
signal $s_i(t) + n(t)$ has an energy $(S_i + N)T$, and the point $\boldsymbol{r}$ representing this signal lies at a distance of $\sqrt{(S_i + N)T}$ from the origin (Fig. 15.6a). The signal vector s_i, the noise vector $\boldsymbol{n}$, and the received vector $\boldsymbol{r}$ are shown in Fig. 15.6a. Because

$$|s_i| = \sqrt{S_i T}, \qquad |\boldsymbol{n}| = \sqrt{NT}, \qquad |\boldsymbol{r}| = \sqrt{(S_i + N)T}$$

it follows that vectors s_i, $\boldsymbol{n}$, and $\boldsymbol{r}$ form a right triangle. Also, $\mathbf{n}$ lies on the sphere of radius $\sqrt{NT}$, centered at s_i. Note that because $\mathbf{n}$ is random, it can lie anywhere on the sphere centered at s_i.*

We have M possible transmitted vectors located inside the big sphere. For each possible s, we can draw a sphere of radius $\sqrt{NT}$ around s. If a received vector $\boldsymbol{r}$ lies on one of the small spheres, the center of that sphere is the transmitted waveform. If we pack the big sphere with M nonoverlapping and nontouching spheres, each of radius $\sqrt{NT}$ (Fig. 15.6b), and use the centers of these M spheres for the transmitted waveforms, we will be able to detect all these M waveforms correctly at the receiver simply by using the maximum-likelihood receiver. The maximum-likelihood receiver looks at the received signal point $\boldsymbol{r}$ and decides that the transmitted signal is that one of the M possible transmitted points that is closest to $\boldsymbol{r}$ (smallest error vector). Every received point $\boldsymbol{r}$ will lie on the surface of one of the M nonoverlapping spheres, and using the maximum-likelihood criterion, the transmitted signal will be chosen correctly as the point lying at the center of the sphere on which $\boldsymbol{r}$ lies.

To compute the number of small spheres that can be packed into the big sphere, we must determine the volume of a sphere of D dimensions.



(b)

(a)

Figure 15.6 (a) Signal space representation of transmitted and received signals and noise signal. (b) Choice of signals for error-free communication.

---

* Because N is the average noise power, the energy over an interval T is $NT - \epsilon$, where $\epsilon \to 0$ as $T \to \infty$. Hence, we can assume that $\mathbf{n}$ lies on the sphere.

Volume of a D-Dimensional Sphere: A D-dimensional sphere is described by the equation

$$x_1^2 + x_2^2 + \cdots + x_D^2 = R^2$$

where R is the radius of the sphere. We can show that the volume $V(R)$ of a sphere of radius R is given by

$$V(R) = R^D V(1) \tag{15.65}$$

where $V(1)$ is the volume of a D-dimensional sphere of unit radius and, thus, is constant. To prove this, we have by definition

$$V(R) = \underset{x_1^2+x_2^2+\cdots+x_D^2 \le R^2}{\iint \cdots \int} dx_1 dx_2 \cdots dx_D$$

Letting $y_j = x_j/R$, we have

$$\begin{aligned} V(R) &= R^D \underset{y_1^2+y_2^2+\cdots+y_D^2 \le 1}{\iint \cdots \int} dy_1 dy_2 \cdots dy_D \\ &= R^D V(1) \end{aligned}$$

Hence, the ratio of the volumes of two spheres of radii $\hat{R}$ and R is

$$\frac{V(\hat{R})}{V(R)} = \left(\frac{\hat{R}}{R}\right)^D$$

A direct consequence of this result is that when D is large, almost all of the volume of the sphere is concentrated at the surface. This is because if $\hat{R}/R < 1$, then $(\hat{R}/R)^D \to 0$ as $D \to \infty$. This ratio approaches zero even if $\hat{R}$ differs from R by a very small amount Δ (Fig. 15.7). This means that no matter how small Δ is, the volume within radius $\hat{R}$ is a negligible fraction of the total volume within radius R if D is large enough. Hence, for a large D, almost all of the volume of a D-dimensional sphere is concentrated at the surface. Such a result sounds strange, but a little reflection will show that it is reasonable. This is because the volume is proportional to the Dth power of the radius. Thus, for large D, a small increase in R can increase the volume tremendously, and all the increase comes from a tiny increase in R near the surface of

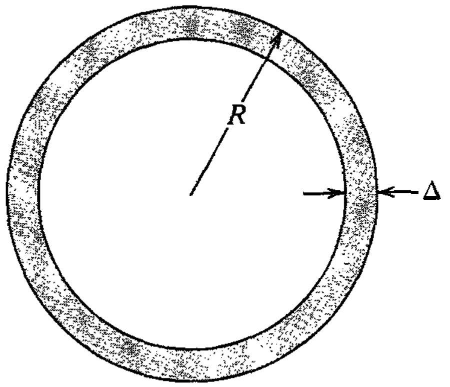


Figure 15.7 Volume of a shell of a D-dimensional hypersphere.

the sphere. This means that most of the volume must be concentrated at the surface.

The number of nonoverlapping spheres of radius $\sqrt{NT}$ that can be packed into a sphere of radius $\sqrt{(S+N)T}$ is bounded by the ratio of the volume of the signal sphere to the volume of the noise sphere. Hence,

$$M \leq \frac{\left[\sqrt{(S+N)T}\right]^{2BT} V(1)}{\left(\sqrt{NT}\right)^{2BT} V(1)} = \left(1+\frac{S}{N}\right)^{BT} \tag{15.66}$$

Each of the M-ary signals carries the information of $\log_2 M$ binary digits. Hence, the transmission of one of the M signals every T seconds is equivalent to the information rate C given by

$$C = \frac{\log M}{T} \leq B \log\left(1+\frac{S}{N}\right) \quad \text{bit/s} \tag{15.67}$$

This equation gives the upper limit of C. To show that we can actually receive error-free information at a rate of $B \log(1 + S/N)$, we use the argument proposed by Shannon.[8] Instead of choosing the M transmitted messages at the centers of nonoverlapping spheres (Fig. 15.6b), Shannon proposed to select the M points randomly located in the signal sphere I_s of radius $\sqrt{ST}$ (Fig. 15.8). Consider one particular transmitted signal s_k. Because the signal energy is assumed to be $\leq S$, point s_k will lie somewhere inside the signal sphere I_s of radius $\sqrt{ST}$. Because all the M signals are picked randomly from this sphere, the probability of finding a signal within a volume ΔV is $M\Delta V/V_s$, where V_s is the volume of I_s. But because for large D all of the volume of the sphere is concentrated at the surface, all M signal points selected randomly would lie near the surface of I_s. Figure 15.8 shows the transmitted signal s_k, the received signal $\boldsymbol{r}$, and the noise $\boldsymbol{n}$. We draw a sphere of radius $\sqrt{NT}$ with $\boldsymbol{r}$ as the center. This sphere intersects the sphere I_s and forms a common lens-shaped region. The signal s_k lies on the surface of both spheres. We shall use a maximum-likelihood receiver. This means that when $\boldsymbol{r}$ is received, we shall make the decision that "s_k was transmitted" provided none of

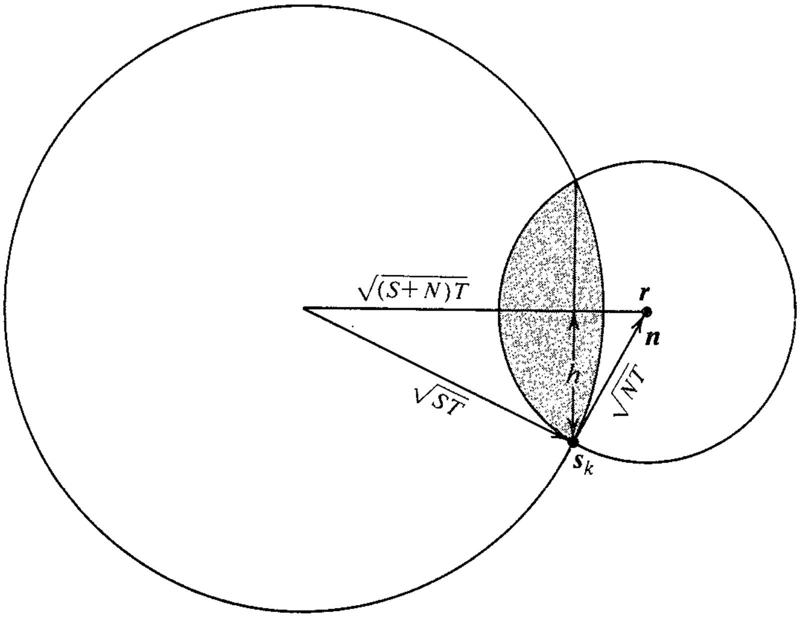


Figure 15.8 Derivation of channel capacity.

the remaining $M - 1$ signal points are closer to $\boldsymbol{r}$ than s_k. The probability of finding any one signal in the lens is V_{lens}/V_s. Hence P_e, the error probability in the detection of s_k when $\boldsymbol{r}$ is received, is

$$\begin{aligned} P_e &= (M-1)\frac{V_{\text{lens}}}{V_s} \\ &< M\frac{V_{\text{lens}}}{V_s} \end{aligned}$$

From Fig. 15.8, we observe that $V_{\text{lens}} < V(h)$, where $V(h)$ is the volume of the D-dimensional sphere of radius h. Because $\boldsymbol{r}$, s_k, and $\boldsymbol{n}$ form a right triangle,

$$h\sqrt{(S+N)T} = \sqrt{(ST)(NT)} \qquad \text{and} \qquad h = \sqrt{\frac{SNT}{S+N}}$$

Hence,

$$V(h) = \left(\frac{SNT}{S+N}\right)^{BT} V(1)$$

Also,

$$V_s = (ST)^{BT} V(1)$$

and

$$P_e < M\left(\frac{N}{S+N}\right)^{BT}$$

If we choose

$$M = \left[k\left(1+\frac{S}{N}\right)\right]^{BT}$$

then

$$P_e < [k]^{BT}$$

If we let $k = 1 - \Delta$, where Δ is a positive number chosen as small as we wish, then

$$P_e \to 0 \qquad \text{as} \qquad BT \to \infty$$

This means that P_e can be made arbitrarily small by increasing T, provided M is chosen arbitrarily close to $(1 + S/N)^{BT}$. Thus,

$$\begin{aligned} C &= \frac{1}{T}\log_2 M \\ &= \left[B\log\left(1+\frac{S}{N}\right) - \epsilon\right] \quad \text{bit/s} \end{aligned} \tag{15.68}$$

where ϵ is a positive number chosen as small as we please. This proves the desired result. A more rigorous derivation of this result can be found in Wozencraft and Jacobs.[9]

Because the M signals are selected randomly from the signal space, they tend to acquire the statistics of white noise[8] (i.e., a white gaussian random process).

Comments on Channel Capacity: According to the result derived in this chapter, theoretically we can communicate error-free up to C bit/s. There are, however, practical difficulties in achieving this rate. In proving the capacity formula, we assumed that communication is effected by signals of duration T. This means we must wait T seconds to accumulate the input data and then encode it by one of the waveforms of duration T. Because the capacity rate is achieved only in the limit as $T \to \infty$, we have to wait a long time at the receiver to get the information. Moreover, because the number of possible messages that can be transmitted over interval T increases exponentially with T, the transmitter and receiver structures increase in complexity beyond imagination as $T \to \infty$.

The channel capacity indicated by Shannon's equation [Eq. (15.68)] is the maximum error-free communication rate achievable on an optimum system without any restrictions (except for bandwidth B, signal power S, and gaussian white channel noise power N). If we have any other restrictions, this maximum rate will not be achieved. For example, if we consider a binary channel (a channel restricted to transmit only binary signals), we will not be able to attain Shannon's rate, even if the channel is optimum. The channel capacity formula [Eq. (15.68)] indicates that the transmission rate is a monotonically increasing function of the signal power S. If we use a binary channel, however, we know that increasing the transmitted power beyond a certain point buys very little advantage (see Fig. 12.14). Hence, on a binary channel, increasing S will not increase the error-free communication rate beyond some value. This does not mean that the channel capacity formula has failed. It simply means that when we have a large amount of power (with a finite bandwidth) available, the binary scheme is not the optimum communication scheme.

One last comment: Shannon's results tell us the upper theoretical limit of error-free communication. But they do not tell us precisely how this can be achieved. To quote the words of Abramson; "[This is one of the problems] which has persisted to mock information theorists since Shannon's original paper in 1948. Despite an enormous amount of effort spent since that time in quest of this Holy Grail of information theory, a *deterministic* method of generating the codes promised by Shannon is still to be found."[4]

15.6 PRACTICAL COMMUNICATION SYSTEMS IN LIGHT OF SHANNON'S EQUATION

It would be instructive to determine the ideal law for the exchange between the SNR and the transmission bandwidth using the channel capacity equation. Consider a message of bandwidth B that is used for modulation (or coding), with the resulting modulated signal of bandwidth B_T. This signal is received at the input of an ideal demodulator with signal and noise powers of S_i and N_i, respectively* (Fig. 15.9). The demodulator output bandwidth is B, and the SNR is S_o/N_o. Because an SNR S/N and a bandwidth B can transmit ideally $B \log(1 + S/N)$ bits of information, the ideal information rates of the signals at the input and the output of the demodulator are $B_T \log(1 + S_i/N_i)$ bits and $B \log(1 + S_o/N_o)$ bits, respectively. Because the demodulator neither creates nor destroys information, the two rates should be equal, that is,

$$B_T \log\left(1 + \frac{S_i}{N_i}\right) = B \log\left(1 + \frac{S_o}{N_o}\right)$$

---

* An additive white gaussian channel noise is assumed.

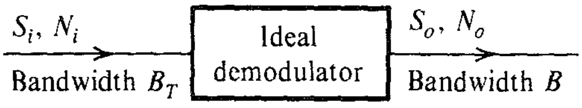


Figure 15.9 Ideal exchange between SNR and bandwidth.

and

$$\left(1+\frac{S_o}{N_o}\right)=\left(1+\frac{S_i}{N_i}\right)^{B_T/B} \tag{15.69a}$$

In practice, for the majority of systems, S_o/N_o as well as $S_i/N_i \gg 1$, and

$$\frac{S_o}{N_o}\simeq\left(\frac{S_i}{N_i}\right)^{B_T/B} \tag{15.69b}$$

Also,

$$\begin{aligned}\frac{S_i}{N_i}&=\frac{S_i}{\mathcal{N}B_T}\\&=\left(\frac{S_i}{\mathcal{N}B}\right)\left(\frac{B}{B_T}\right)=\frac{B}{B_T}\gamma \qquad \gamma=\frac{S_i}{\mathcal{N}B}\end{aligned}$$

Hence, Eqs. (15.69) become

$$\frac{S_o}{N_o}=\left(1+\frac{\gamma}{B_T/B}\right)^{B_T/B}-1 \tag{15.70a}$$

$$\simeq\left(\frac{\gamma}{B_T/B}\right)^{B_T/B} \tag{15.70b}$$

Equations (15.69) and (15.70) give the ideal law of exchange between the SNR and the bandwidth. The output SNR S_o/N_o is plotted in Fig. 15.10 as a function of γ for various values of B_T/B.

The output SNR increases exponentially with the bandwidth expansion factor B_T/B. This means that to maintain a given output SNR, the transmitted signal power can be reduced exponentially with the bandwidth expansion factor. Thus, for a small increase in bandwidth, we can reduce the transmitted power considerably. On the other hand, for a small reduction in bandwidth, we need to increase the transmitted power considerably. Hence, in practice, the trade is in the sense of reducing the transmitted power at the cost of increased transmission bandwidth and rarely the other way.

Let us now investigate various systems studied thus far and see how they fare in comparison with the ideal system.

AM

For baseband and SSB-SC systems, $B_T/B = 1$, and Eqs. (15.70) yield

$$\frac{S_o}{N_o}=\gamma \tag{15.71}$$

which is exactly the performance of these systems [see Eqs. (12.3) and (12.12)]. For DSB-SC, $B_T/B = 2$, and Eqs. (15.70) predict

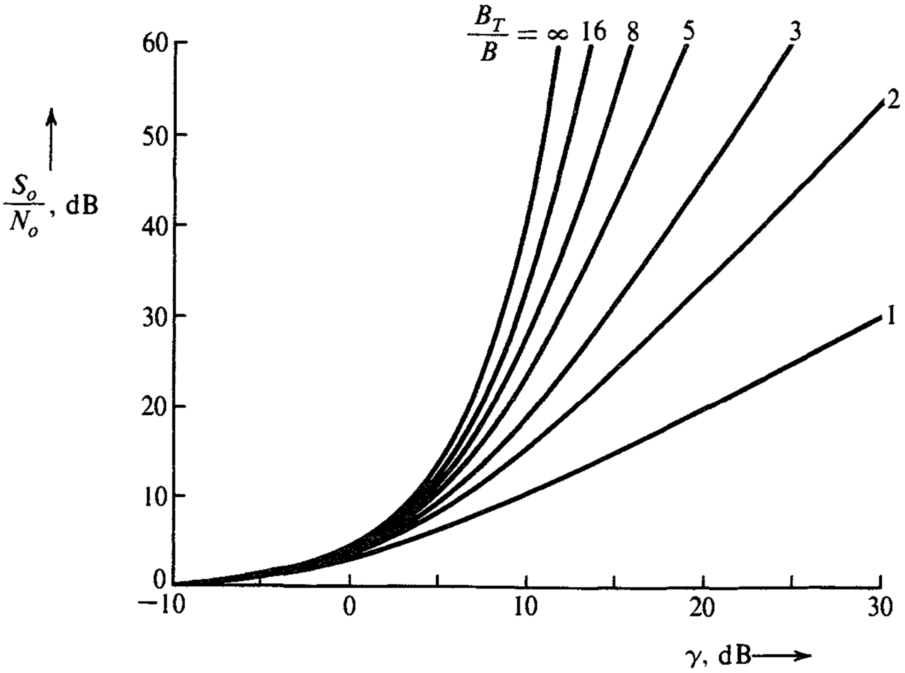


Figure 15.10 Ideal behavior of SNR vs. γ for various ratios of B_T/B.

$$\frac{S_o}{N_o} \simeq \frac{\gamma^2}{4} \tag{15.72}$$

Thus DSB-SC, for which $S_o/N_o = \gamma$, falls short of ideal performance. If, however, we consider the fact that quadrature multiplexing can be used to transmit two DSB signals simultaneously, with the effective bandwidth per signal as B rather than $2B$, DSB-SC has ideal performance. Because for AM quadrature multiplexing is not used,* AM performance [Eq. (12.14)] falls considerably short of the ideal performance [Eq. (15.72)].

The ideal performance of the baseband and SSB-SC or DSB-SC systems is really an empty boast, because these systems do not exchange SNR for bandwidth.

FM

For FM, we have [Eq. (12.37)]

$$\frac{S_o}{N_o} = 3\beta^2\gamma\left(\frac{\overline{m^2}}{m_p^2}\right)$$

Because $B_T/B = 2(\beta + 1)$,

$$\frac{S_o}{N_o} = 3\left[\frac{1}{2}\left(\frac{B_T}{B}\right) - 1\right]^2\left(\frac{\overline{m^2}}{m_p^2}\right)\gamma \tag{15.73}$$

Figure 15.11 shows the plots of S_o/N_o for $B_T/B = 8$ and 20, assuming $m_p^2/\overline{m^2} = 2$. The ideal S_o/N_o plots for $B_T/B = 8$ and 20 are also shown in the figure for comparison.

---

* Because this requires a phase reference at the receiver, it defeats the purpose of AM.

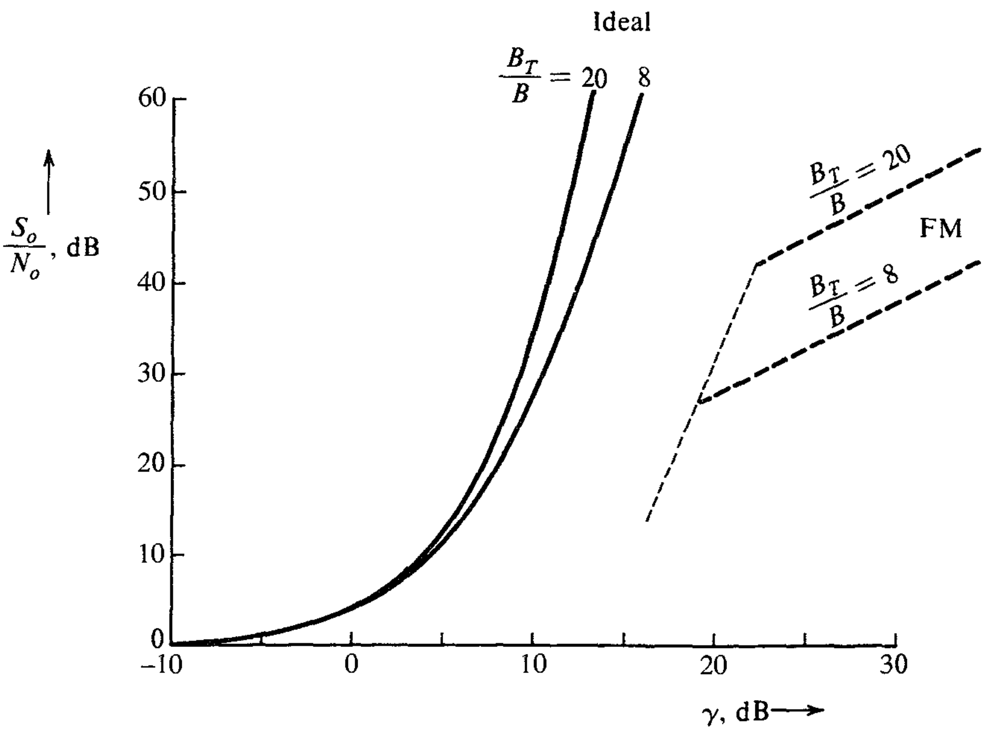


Figure 15.11 Comparison of ideal behavior with FM system behavior.

FM performance falls far below ideal performance. Even if a gain of 13 dB resulting from preemphasis and deemphasis is added to these plots, FM is several decibels inferior to ideal curves. The comparison between FM and the ideal system gets progressively worse as γ increases. If we observe the behavior of FM at the threshold, however, it does not fare as badly. It can be shown that for FM, when optimum demodulation (phase-locked loop) is used, the behavior of the SNR at the threshold (the dashed line* in Fig. 15.11) is close to ideal.[10]

PCM

As seen earlier, M-ary PCM shows a saturation effect unless we go to higher values of M as γ increases. If the message signal is quantized in L levels, then each sample can be encoded by $\log_M L$ number of M-ary pulses. If B is the bandwidth of the message signal, we need to transmit $2B$ samples per second. Consequently, R_M, the number of M-ary pulses per second, is

$$R_M = 2B \log_M L$$

Also, the transmission bandwith B_T is half the number of (M-ary) pulses per second. Hence,

$$B_T = \frac{R_M}{2} = B \log_M L \tag{15.74a}$$

From Eq. (13.51a), the power S_i is found as

$$S_i = \frac{M^2 - 1}{3} E_p R_M \tag{15.74b}$$

---

* The dashed threshold line shown in Fig. 15.11 is for a frequency discriminator. For optimum demodulation using a phase-locked loop, the threshold line is shifted left by 3 to 5 dB.

Also,

$$N_i = \mathcal{N} B_T = \frac{\mathcal{N} R_M}{2} \tag{15.75}$$

Each of the M-ary pulses carries the information of $\log_2 M$ bits, and we are transmitting $2B \log_M L$ number of M-ary pulses per second. Hence, we are transmitting information at a rate of R_b bits, where

$$\begin{aligned} R_b &= (2B \log_M L)(\log_2 M) \\ &= 2B_T \log_2 M \\ &= B_T \log_2 M^2 \quad \text{bit/s} \end{aligned}$$

Substitution of Eqs. (15.74b) and (15.75) into this equation yields

$$R_b = B_T \log_2 \left(1 + \frac{3\mathcal{N}}{2E_p} \frac{S_i}{N_i}\right) \quad \text{bit/s} \tag{15.76}$$

We are transmitting the information equivalent of R_b binary digits per second over the M-ary PCM channel. The reception is not error-free, however. The pulses are detected with an error probability P_{eM} given in Eq. (13.52b). If P_{eM} is on the order of 10^{-6}, we could consider the reception to be essentially error-free. From Eq. (13.52b),

$$P_{eM} \simeq 2Q\left(\sqrt{\frac{2E_p}{\mathcal{N}}}\right) = 10^{-6} \qquad M \gg 1$$

This gives

$$\frac{2E_p}{\mathcal{N}} = 24$$

Substitution of this value in Eq. (15.76) gives

$$R_b = B_T \log_2 \left(1 + \frac{1}{8} \frac{S_i}{N_i}\right) \quad \text{bit/s} \tag{15.77}$$

Thus, over a channel of bandwith B_T with an SNR of S_i/N_i, a PCM system can transmit information at a rate of R_b in Eq. (15.77). The ideal channel with bandwidth B_T and SNR S_i/N_i transmits information at a rate of C bit/s, where

$$C = B_T \log_2 \left(1 + \frac{S_i}{N_i}\right) \quad \text{bit/s} \tag{15.78}$$

It follows that PCM uses roughly eight times (9 dB) as much power as the ideal system. This performance is still much superior to that of FM. Figure 15.12 shows R_b/B_T as a function of S_i/N_i. For the ideal system

$$\frac{R_b}{B_T} = \frac{C}{B_T} = \log_2 \left(1 + \frac{S_i}{N_i}\right)$$

PCM at the threshold is 9 dB inferior to the ideal curve.

When PCM is in saturation, the detection error probability approaches 0. Each M-ary pulse transmits $\log_2 M$ bits, and there are $2B_T$ pulses per second. Hence,

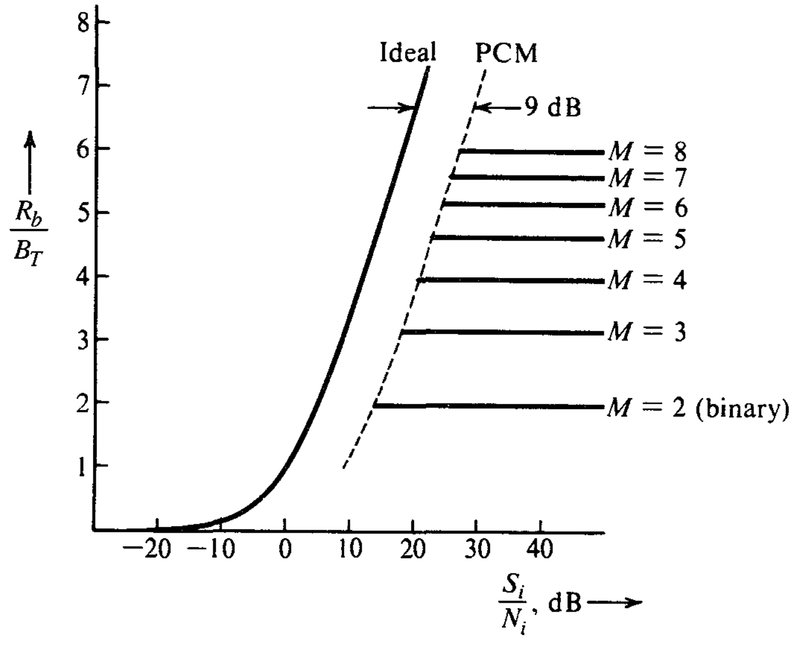


Figure 15.12 Comparison of ideal system behavior with that of PCM.

$$R_b = 2B_T \log_2 M$$

or

$$\frac{R_b}{B_T} = 2 \log_2 M$$

This is clearly seen in Fig. 15.12.

Orthogonal Signaling

We have already shown that [see Eq. (13.58)] for M-ary orthogonal signaling, the error-free communication rate is

$$R_b \leq 1.44 \frac{S_i}{\mathcal{N}} \quad \text{bit/s} \tag{15.79}$$

We have shown in Eq. (15.64) that this is precisely the rate of error-free communication over an an ideal channel with infinite bandwidth. Therefore, as $M \to \infty$, the bandwidth of an M-ary scheme approaches ∞, and its rate of communication approaches that of an ideal channel.

REFERENCES

1. C. E. Shannon, "Mathematical Theory of Communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, July 1948; pp. 623–656, Oct. 1948.
2. R. V. L. Hartley, "Transmission of Information," *Bell Syst. Tech. J.*, vol. 7, pp. 535–563, July 1928.
3. H. Nyquist, "Certain Factors Affecting Telegraph Speed," *Bell Syst. Tech. J.*, vol. 3, pp. 324–346, April 1924.
4. N. Abramson, *Information Theory and Coding*, McGraw-Hill, New York, 1963.
5. R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

6. D. A. Huffman, "A Method for Construction of Minimum Redundancy Codes," *Proc. IRE*, vol. 40, pp. 1098–1101, Sept. 1952.
7. R. W. Hamming, *Coding and Information Theory*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1986.
8. C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.
9. J. M. Wozencraft and I. A. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965, chap. 5.
10. A. J. Viterbi, *Principles of Coherent Communication*, McGraw-Hill, New York, 1966.

PROBLEMS

15.1-1 A message source generates one of four messages randomly every microsecond. The probabilities of these messages are 0.4, 0.3, 0.2, and 0.1. Each emitted message is independent of the other messages in the sequence.

(a) What is the source entropy?

(b) What is the rate of information generated by this source (in bits per second)?

15.1-2 A television picture is composed of approximately 300,000 basic picture elements (about 600 picture elements in a horizontal line and 500 horizontal lines per frame). Each of these elements can assume 10 distinguishable brightness levels (such as black and shades of gray) with equal probability. Find the information content of a television picture frame.

15.1-3 A radio announcer describes a television picture orally in 1000 words out of his vocabulary of 10,000 words. Assume that each of the 10,000 words in his vocabulary is equally likely to occur in the description of this picture (a crude approximation, but good enough to give an idea). Determine the amount of information broadcast by the announcer in describing the picture. Would you say the announcer can do justice to the picture in 1000 words? Is the old adage "a picture is worth a thousand words" an exaggeration or an underrating of the reality? Use data in Prob. 15.1-2 to estimate the information of a picture.

15.1-4 From the Old North Church Tower in Boston, Paul Revere's friend was to show him one lantern if the British army began advancing overland and two lanterns if they chose to cross the bay in boats.

(a) Assume Revere had no way of guessing ahead of time what route the British might choose. How much information did he receive when he saw two lanterns?

(b) What if Revere were 90% sure the British would march overland? Then, how much information would the two lanterns have conveyed?

15.1-5 Estimate the information per letter in the English language by various methods, assuming each character is independent of the others. (This is not true, but is good enough to get a rough idea.)

(a) In the first method, assume that all 27 characters (26 letters and a space) are equiprobable. This is a gross approximation, but is good for a quick answer.

(b) In the second method, use the table of probabilities of various characters (Table P15.1-5).

(c) Lastly, use Zipf's law relating the word rank to its probability. In English prose, if we order words according to the frequency of usage so that the most frequently used word (*the*) is word number 1 (rank 1), the next most probable word (*of*) is number 2 (rank 2), and so on, then empirically it is found that $P(r)$, the probability of the rth word (rank r) is very nearly

$$P(r) = \frac{0.1}{r}$$

Using Zipf's law, compute the entropy per word. Assume that there are 8727 words. The reason for this number is that the probabilities $P(r)$ sum to 1 for r from 1 to 8727. Zipf's law, surprisingly, gives reasonably good results. Assuming there are 5.5 letters (including space) on average per word, determine the entropy or information per letter.

Table P15.1.5
Probability of occurrence of letters in the English language

Letter	Probability	$-\log P_i$	Letter	Probability	$-\log P_i$
Space	0.187	2.46	M	0.02075	5.60
E	0.1073	3.22	U	0.02010	5.64
T	0.0856	3.84	G	0.01633	5.94
A	0.0668	3.90	Y	0.01623	5.95
O	0.0654	3.94	P	0.01623	5.95
N	0.0581	4.11	W	0.01620	6.32
R	0.0559	4.16	B	0.01179	6.42
I	0.0519	4.27	V	0.00752	7.06
S	0.0499	4.33	K	0.00344	8.20
H	0.04305	4.54	X	0.00136	9.54
D	0.03100	5.02	J	0.00108	9.85
L	0.02775	5.17	Q	0.00099	9.98
F	0.02395	5.38	Z	0.00063	10.63
C	0.02260	5.45			

15.2-1 A source emits seven messages with probabilities 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, and 1/64, respectively. Find the entropy of the source. Obtain the compact binary code and find the average length of the code word. Determine the efficiency and the redundancy of the code.

15.2-2 A source emits seven messages with probabilities 1/3, 1/3, 1/9, 1/9, 1/27, 1/27, and 1/27, respectively. Find the entropy of the source. Obtain the compact 3-ary code and find the average length of the code word. Determine the efficiency and the redundancy of the code.

15.2-3 A source emits one of four messages randomly every 1 microsecond. The probabilities of these messages are 0.5, 0.3, 0.1, and 0.1. Messages are generated independently.

(a) What is the source entropy?

(b) Obtain a compact binary code and determine the average length of the code word, the efficiency, and the redundancy of the code.

(c) Repeat part **(b)** for a compact ternary code.

15.2-4 For the messages in Prob. 15.2-1, obtain the compact 3-ary code and find the average length of the code word. Determine the efficiency and the redundancy of this code.

15.2-5 For the messages in Prob. 15.2-2, obtain the compact binary code and find the average length of the code word. Determine the efficiency and the redundancy of this code.

15.2-6 A source emits three equiprobable messages randomly and independently.

(a) Find the source entropy.

(b) Find a compact ternary code, the average length of the code word, the code efficiency, and the redundancy.

(c) Repeat part **(b)** for a binary code.

(d) To improve the efficiency of a binary code, we now code the second extension of the source. Find a compact binary code, the average length of the code word, the code efficiency, and the redundancy.

15.4-1 A binary channel matrix is given by

$$\begin{array}{cccc} & & \multicolumn{2}{c}{\text{Outputs}} \\ & & y_1 & y_2 \\ \text{Inputs} & \begin{array}{c} x_1 \\ \\ x_2 \end{array} & \left[\begin{array}{c} \frac{2}{3} \\ \\ \frac{1}{10} \end{array}\right. & \left.\begin{array}{c} \frac{1}{3} \\ \\ \frac{9}{10} \end{array}\right] \end{array}$$

This means $P_{y|x}(y_1|x_1) = 2/3$, $P_{y|x}(y_2|x_1) = 1/3$, etc. You are also given that $P_x(x_1) = 1/3$ and $P_x(x_2) = 2/3$. Determine $H(x)$, $H(x|y)$, $H(y)$, $H(y|x)$, and $I(x; y)$.

15.4-2 For the ternary channel in Fig. P15.4-2, $P_x(x_1) = P$, $P_x(x_2) = P_x(x_3) = Q$. (*Note*: $P+2Q = 1$.)

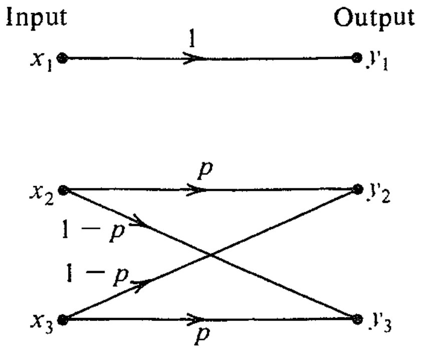


Figure P15.4-2

(a) Determine $H(x)$, $H(x|y)$, $H(y)$, and $I(x; y)$.

(b) Show that the channel capacity C_s is given by

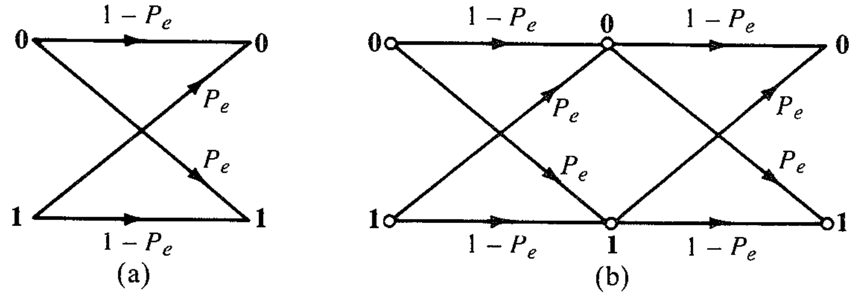
$$C_s = \log\left(\frac{\beta+2}{\beta}\right)$$

where $\beta = 2^{-[p\log p+(1-p)\log(1-p)]}$.

15.4-3 Consider the binary symmetric channel shown in Fig. P15.4-3a. The channel matrix is given by

$$M = \begin{bmatrix} 1-P_e & P_e \\ P_e & 1-P_e \end{bmatrix}$$

Figure P15.4-3b shows a cascade of two such BSCs.



(a) (b)

Figure P15.4-3

(a) Determine the channel matrix for the cascaded channel in Fig. P15.4-3b. Show that this matrix is M^2.

(b) If the two BSC channels in Fig. P15.4-3b have error probabilities P_{e1} and P_{e2}, with channel matrices M_1 amd M_2, respectively, show that the channel matrix of the cascade of these two channels is $M_1 M_2$.

(c) Using the results in part (b), show that the channel matrix for the cascade of k identical BSCs each with channel matrix M is M^k. Verify your answer for $n = 3$ by confirming the results in Example 10.7.

(d) Use the result in part (c) to determine the channel capacity for a cascade of k identical BSC channels each with error probability P_e.

15.4-4 In data communication using error detection code, as soon as an error is detected, an automatic request for retransmission (ARQ) enables retransmission of the data in error. In such a channel the data in error is erased. Hence, there is an erase probability p, but the probability of error is zero. Such a channel, known as a **binary erasure channel (BEC)**, can be modeled as shown in Fig. P15.4-4. Determine $H(\mathrm{x})$, $H(\mathrm{x}|\mathrm{y})$, and $I(\mathrm{x};\mathrm{y})$ assuming the two transmitted messages equiprobable.

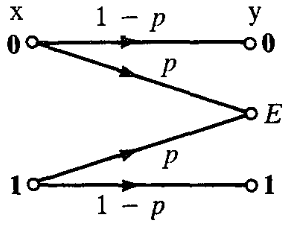


Figure P15.4-4

15.4-5 A cascade of two channels is shown in Fig. P15.4-5. The symbols at the source, at the output of the first channel, and at the output of the second channel are denoted by x, y, and z. Show that

$$H(\mathrm{x}|\mathrm{z}) \geq H(\mathrm{x}|\mathrm{y})$$

and

$$I(\mathrm{x};\mathrm{y}) \geq I(\mathrm{x};\mathrm{z})$$

This shows that the information that can be transmitted over a cascaded channel can be no greater than that transmitted over one link. In effect, information channels tend to leak information. *Hint:* For a cascaded channel, observe that

$$P(\mathrm{z}_k|\mathrm{y}_j, \mathrm{x}_i) = P(\mathrm{z}_k|\mathrm{y}_j)$$

Hence, by Bayes' rule,

$$P(\mathrm{x}_k|\mathrm{y}_j, \mathrm{z}_k) = P(\mathrm{x}_i|\mathrm{y}_j)$$

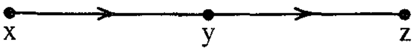


Figure P15.4-5

15.5-1 For a continuous random variable x constrained to a peak magnitude M $(-M < x < M)$, show that the entropy is maximum when x is uniformly distributed in the range $(-M, M)$ and has zero probability density outside this range. Show that the maximum entropy is given by $\log 2M$.

15.5-2 For a continuous random variable x constrained to only positive values $0 < x < \infty$ and a mean value A, show that the entropy is maximum when

$$P_x(x) = \frac{1}{A} e^{-x/A} u(x)$$

Show that the corresponding entropy is

$$H(\mathrm{x}) = \log eA$$

15.5-3 A television transmission requires 30 frames of 300,000 picture elements each to be transmitted per second. Using the data in Prob. 15.1-2, estimate the theoretical bandwidth of the AWGN channel if the SNR at the receiver is required to be at least 50 dB.

15.5-4 Show that the channel capacity of a band-limited channel disturbed by a colored gaussian noise under the constraint of a given signal power is

$$C = B \log [S_s(\omega) + S_n(\omega)] - \int_{f_1}^{f_2} \log S_n(\omega)\, df \qquad df = \frac{d\omega}{2\pi}$$

where B is the channel bandwidth (in hertz) over the frequency range (f_1, f_2) $(f_2 - f_1 = B)$. $S_s(\omega)$ and $S_n(\omega)$ are the signal and the noise power densities, respectively. Show that this maximum rate of information transmission is attained if the desired signal is gaussian and its PSD satisfies the condition

$$S_s(\omega) + S_n(\omega) = \alpha \quad \text{(a constant)}$$

Hint: Consider a narrow-band Δf in the range (f_1, f_2). The maximum rate of transmission over this band is given by

$$\Delta f \log \left[\frac{S_s(\omega)\Delta f + S_n(\omega)\Delta f}{S_n(\omega)\Delta f}\right] = \Delta f \log \left[\frac{S_s(\omega) + S_n(\omega)}{S_n(\omega)}\right]$$

provided the signal over this band is gaussian. The rate of transmission over the entire band is given by

$$\int_{f_1}^{f_2} \log \left[\frac{S_s(\omega) + S_n(\omega)}{S_n(\omega)}\right] df$$

Now maximize this under the constraint

$$2\int_{f_1}^{f_2} S_s(\omega)\, df = S$$

15.5-5 Using the results in Prob. 15.5-4, show that the worst kind of gaussian noise is a white gaussian noise that is constrained to a given mean square value. *Hint:* Use the expression for channel capacity in Prob. 15.5-4. The first term in this expression is a constant. Now show that the second term attains a maximum value when $S_n(\omega)$ is a constant under the constraint

$$2\int_{f_1}^{f_2} S_n(\omega)\, df = N$$

16 ERROR CORRECTING CODES

As seen from the discussion in Chapter 15, the key to realizing error-free communication is the use of appropriate redundancy. The addition of a single parity-check digit to detect an odd number of errors is a good example. Since Shannon's pioneering paper, a great deal of work has been carried out in the area of error correcting codes.

16.1 INTRODUCTION

In this chapter we shall discuss two important types of codes: block codes and convolutional codes. The information coming from the data or message source will be assumed to be in binary form (a sequence of binary digits).

In **block codes**, a block of k data digits is encoded by a code word of n digits ($n > k$). For each sequence of k data digits, there is a distinct code word of n digits. In **convolutional**, or **recurrent codes**, the coded sequence of n digits depends not only on the k data digits but also on the previous $N-1$ data digits ($N > 1$). Hence, the coded sequence for a certain k data digits is not unique but depends on $N-1$ earlier data digits. In block codes, k data digits are accumulated and then encoded into an n-digit code word. In convolutional codes, the coding is done on a continuous, or running, basis rather than by accumulating k data digits.

If k data digits are transmitted by a code word of n digits, the number of check digits is $m = n - k$. The **code efficiency** (also known as the **code rate**) is k/n. Such a code is known as an (n, k) code. Data digits $(d_1, d_2, \ldots, d_k)$ are a k-tuple, and, hence, this is a k-dimensional vector $\boldsymbol{d}$. Similarly, a code word $(c_1, c_2, \ldots, c_n)$ is an n-dimensional vector $\boldsymbol{c}$. As a preliminary, we shall determine the minimum number of check digits required to detect or correct t number of errors in an (n, k) code.

A total of 2^n code words (or vertices of an n-dimensional hypercube) is available to assign to 2^k data words. Suppose we wish to find a code that will correct up to t wrong digits. In this case, if we transmit a data word $\boldsymbol{d}_j$ by one of the code words (or vertices) $\boldsymbol{c}_j$, then because of channel errors the received word will not be $\boldsymbol{c}_j$ but will be $\boldsymbol{c}'_j$. If the channel noise causes errors in t or less digits, then c_j' will lie somewhere in the Hamming sphere of radius t centered at c_j. If the code is to correct up to t errors, then the code must have the property that all of the Hamming spheres of radius t centered at the code words are nonoverlapping. This means we may not use vertices (or words) that are within a Hamming distance of t from any code word. If a received word lies within a Hamming sphere of radius t centered at c_j, then we decide that the transmitted code word was c_j. This scheme is capable of correcting up to t errors, and d_{min}, the minimum distance between t error correcting code words, is

$$d_{\text{min}} = 2t + 1 \tag{16.1}$$

Next, in order to find a relationship between n and k, we observe that 2^n vertices, or words, are available for 2^k data words, and $2^n - 2^k$ are redundant vertices. How many vertices, or words, can lie within a Hamming sphere of radius t? The number of sequences (of n digits) that differ from a given sequence by j digits is the number of the combination of n things taken j at a time and is given by $\binom{n}{j}$ [see Eq. (10.16)]. Hence, the number of ways in which up to t errors can occur is given by $\sum_{j=1}^{t} \binom{n}{j}$. Thus for each code word, we must leave $\sum_{j=1}^{t} \binom{n}{j}$ number of words unused. Because we have 2^k code words, we must leave $2^k \sum_{j=1}^{t} \binom{n}{j}$ words unused. Hence, the total number of words must be at least

$$2^k + 2^k \sum_{j=1}^{t} \binom{n}{j} = 2^k \sum_{j=0}^{t} \binom{n}{j}$$

But the total number of words, or vertices, available is 2^n. Hence,

$$2^n \geq 2^k \sum_{j=0}^{t} \binom{n}{j}$$

or

$$2^{n-k} \geq \sum_{j=0}^{t} \binom{n}{j} \tag{16.2a}$$

Observe that $n - k = m$ is the number of check digits. Hence, Eq. (16.2a) can be expressed as

$$2^m \geq \sum_{j=0}^{t} \binom{n}{j} \tag{16.2b}$$

This is known as the **Hamming bound**. It should also be remembered that the Hamming bound is a necessary but not a sufficient condition in general. However, for single-error correcting codes, it is a necessary and sufficient condition. If some m satisfies the Hamming bound, it does not necessarily mean that a t-error correcting code of n digits can be constructed. Table 16.1 shows some examples of error correction codes and their efficiencies.

A code for which the inequalities in Eqs. (16.2) become equalities is known as the **perfect code**. In such a code the Hamming spheres (about all the code words) are not only nonoverlapping but they exhaust all the 2^n vertices, leaving no vertex outside some sphere. An e-error correcting perfect code satisfies the condition that every possible sequence is at

Table 16.1
Some examples of error correcting codes

	n	k	Code	Code Efficiency (or Code Rate)
Single-error correcting, $t = 1$	3	1	(3, 1)	0.33
	4	1	(4, 1)	0.25
Minimum code separation 3	5	2	(5, 2)	0.4
	6	3	(6, 3)	0.5
	7	4	(7, 4)	0.57
	15	11	(15, 11)	0.73
	31	26	(31, 26)	0.838
Double-error correcting, $t = 2$	10	4	(10, 4)	0.4
	15	8	(15, 8)	0.533
Minimum code separation 5				
Triple-error correcting, $t = 3$	10	2	(10, 2)	0.2
	15	5	(15, 5)	0.33
Minimum code separation 7	23	12	(23, 12)	0.52

a distance at most e from some code word. Perfect codes exist in only a comparatively few cases. Binary, single-error correcting, perfect codes are called **Hamming codes.**

For a Hamming code, $t = 1$ and $d_{\min} = 3$, and from Eq. (16.2b) we have

$$2^m = \sum_{j=0}^{1} \binom{n}{j} = 1 + n \tag{16.3}$$

and

$$n = 2^m - 1$$

Another way of correcting errors is to design a code to detect (not to correct) up to t errors. When the receiver detects an error, it requests retransmission. Because error detection requires fewer check digits, these codes operate at a higher efficiency.

To detect t errors, code words need to be separated by a Hamming distance of not more than $t + 1$. Suppose a transmitted code word c_j has α number of errors ($\alpha \leq t$). Then the received codeword c_j' is at a distance of α from c_j. Because $\alpha \leq t$, c_j' can never be any other valid code word, because all code words are separated by at least $t + 1$. Thus, the reception of c_j' immediately indicates that an error has been made.

Thus, the minimum distance $d_{\min}$ between t error detecting code words is

$$d_{\min} = t + 1$$

In presenting the theory, we shall use modulo-2 addition, defined in Chapter 7:

$$\mathbf{1 \oplus 1 = 0 \oplus 0 = 0}$$
$$\mathbf{0 \oplus 1 = 1 \oplus 0 = 1}$$

Note that the modulo-2 sum of any binary digit with itself is always zero. All the additions in the mathematical development of binary codes presented henceforth are modulo-2.

16.2 LINEAR BLOCK CODES

A code word consists of n digits $c_1, c_2, \ldots, c_n$, and a data word consists of k digits $d_1, d_2, \ldots, d_k$. Because the code word and the data word are an n-tuple and a k-tuple, respectively, they are n- and k-dimensional vectors. We shall use row matrices to represent these words,

$$\boldsymbol{c} = (c_1, c_2, \ldots, c_n), \qquad \boldsymbol{d} = (d_1, d_2, \ldots, d_k)$$

For the general case of linear block codes, all the n digits of $\boldsymbol{c}$ are formed by linear combinations (modulo-2 additions) of k data digits. A special case where $c_1 = d_1, c_2 = d_2, \ldots, c_k = d_k$ and the remaining digits from c_{k+1} to c_n are linear combinations of $d_1, d_2, \ldots, d_k$ is known as a **systematic code.*** Thus in a systematic code, the first k digits of a code word are the data digits and the last $m = n - k$ digits are the **parity-check digits**, formed by linear combinations of data digits $d_1, d_2, \ldots, d_k$:

$$\begin{aligned} c_1 &= d_1 \\ c_2 &= d_2 \\ &\vdots \\ c_k &= d_k \\ c_{k+1} &= h_{11}d_1 \oplus h_{12}d_2 \oplus \cdots \oplus h_{1k}d_k \\ c_{k+2} &= h_{21}d_1 \oplus h_{22}d_2 \oplus \cdots \oplus h_{2k}d_k \\ &\vdots \\ c_n &= h_{m1}d_1 \oplus h_{m2}d_2 \oplus \cdots \oplus h_{mk}d_k \end{aligned} \tag{16.4a}$$

or

$$\boldsymbol{c} = \boldsymbol{dG} \tag{16.4b}$$

where

$$\boldsymbol{G} = \left[\underbrace{\begin{matrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \end{matrix}}_{\boldsymbol{I}_k(k \times k)} \qquad \underbrace{\begin{matrix} h_{11} & h_{21} & \cdots & h_{m1} \\ h_{12} & h_{22} & \cdots & h_{m2} \\ & & \vdots & \\ h_{1k} & h_{2k} & \cdots & h_{mk} \end{matrix}}_{\boldsymbol{P}(k \times m)} \right] \tag{16.5}$$

The $k \times n$ matrix $\boldsymbol{G}$ is called the **generator matrix**, which can be partitioned into a $k \times k$ identity matrix $\boldsymbol{I}_k$ and a $k \times m$ matrix $\boldsymbol{P}$. All the elements of $\boldsymbol{P}$ are either 0 or 1. The code word can be expressed as

$$\begin{aligned} \boldsymbol{c} &= \boldsymbol{dG} \\ &= \boldsymbol{d}[\boldsymbol{I}_k, \boldsymbol{P}] \\ &= [\boldsymbol{d}, \boldsymbol{dP}] \\ &= [\boldsymbol{d}, \boldsymbol{c}_p] \end{aligned} \tag{16.6}$$

---

* It can be shown that the performance of systematic block codes is identical to that of nonsystematic block codes.

where

$$c_p = \boldsymbol{d}\boldsymbol{P} \tag{16.7}$$

Thus, knowing the data digits, we can calculate the check digits from Eq. (16.7).

EXAMPLE 16.1 For a (6, 3) code, the generator matrix $\boldsymbol{G}$ is

$$\boldsymbol{G} = \left[\underbrace{\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}}_{I_k} \quad \underbrace{\begin{matrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{matrix}}_{P} \right]$$

For all eight possible data words, find the corresponding code words, and verify that this code is a single-error correcting code.

Table 16.2 shows the eight data words and the corresponding code words found from $\boldsymbol{c} = \boldsymbol{d}\boldsymbol{G}$.

Table 16.2

Data Word *d*	Code Word *c*
111	111000
110	110110
101	101011
100	100101
011	011101
010	010011
001	001110
000	000000

Note that the distance between any two code words is at least 3. Hence, the code can correct at least one error. Figure 16.1 shows a possible encoder for this code, using a three-digit shift register and three modulo-2 adders.

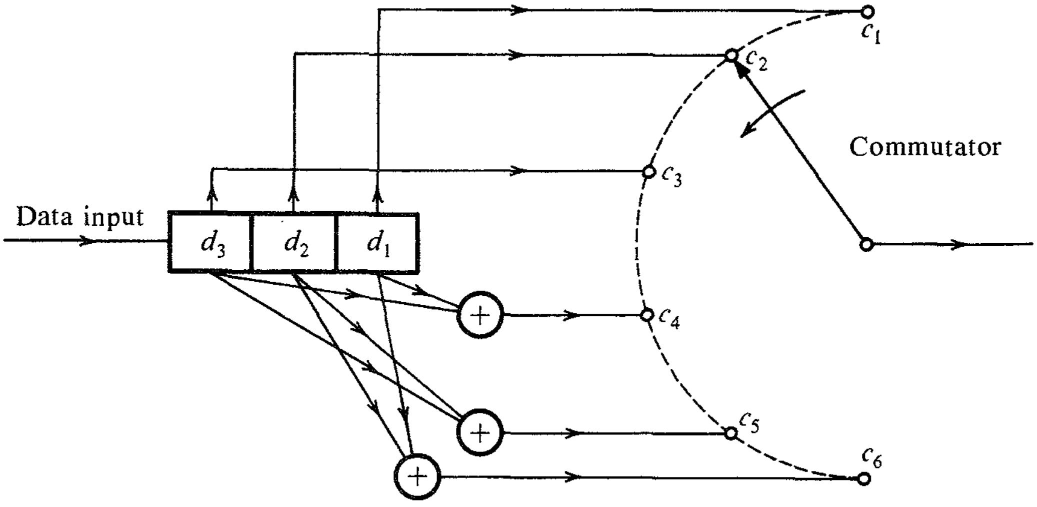


Figure 16.1 Encoder for linear block codes.

Decoding

Let us consider some code word properties that could be utilized for the purpose of decoding. From Eq. (16.7) and the fact that the modulo-2 sum of any sequence with itself is zero, we get

$$\boldsymbol{d}\boldsymbol{P} \oplus \boldsymbol{c}_p = \underbrace{[\boldsymbol{d} \quad \boldsymbol{c}_p]}_{\boldsymbol{c}} \begin{bmatrix} \boldsymbol{P} \\ \boldsymbol{I}_m \end{bmatrix} = 0 \tag{16.8}$$

where $\boldsymbol{I}_m$ is the identity matrix of order $m \times m$ $(m = n - k)$. Thus,

$$\boldsymbol{c}\boldsymbol{H}^T = 0 \tag{16.9a}$$

where

$$\boldsymbol{H}^T = \begin{bmatrix} \boldsymbol{P} \\ \boldsymbol{I}_m \end{bmatrix} \tag{16.9b}$$

and its transpose

$$\boldsymbol{H} = [\boldsymbol{P}^T \quad \boldsymbol{I}_m] \tag{16.9c}$$

is called the **parity-check matrix**. Every code word must satisfy Eq. (16.9a). This is our clue to decoding. Consider the received word $\boldsymbol{r}$. Because of possible errors caused by channel noise, $\boldsymbol{r}$ in general differs from the transmitted code word $\boldsymbol{c}$,

$$\boldsymbol{r} = \boldsymbol{c} \oplus \boldsymbol{e}$$

where the error word (or error vector) $\boldsymbol{e}$, is also a row vector of n elements. For example, if the data word **100** in Example 16.1 is transmitted as a code word **100101** (see Table 16.2), and the channel noise causes a detection error in the third digit, then

$$\boldsymbol{r} = \mathbf{101101}$$

$$\boldsymbol{c} = \mathbf{100101}$$

and

$$\boldsymbol{e} = \mathbf{001000}$$

Thus, an element **1** in $\boldsymbol{e}$ indicates an error in the corresponding position, and **0** indicates no error. The Hamming distance between $\boldsymbol{r}$ and $\boldsymbol{c}$ is simply the number of **1**'s in $\boldsymbol{e}$.

Suppose the transmitted code word is $\boldsymbol{c}_i$ and the channel noise causes an error $\boldsymbol{e}_i$, making the received word $\boldsymbol{r} = \boldsymbol{c}_i + \boldsymbol{e}_i$. If there were no errors, that is, if $\boldsymbol{e}_i = \mathbf{000000}$, then $\boldsymbol{r}\boldsymbol{H}^T = 0$. But because of possible channel errors, $\boldsymbol{r}\boldsymbol{H}^T$ is in general a nonzero row vector $\boldsymbol{s}$, called the **syndrome**,

$$\begin{aligned} \boldsymbol{s} &= \boldsymbol{r}\boldsymbol{H}^T && (16.10a) \\ &= (\boldsymbol{c}_i \oplus \boldsymbol{e}_i)\boldsymbol{H}^T \\ &= \boldsymbol{c}_i\boldsymbol{H}^T \oplus \boldsymbol{e}_i\boldsymbol{H}^T \\ &= \boldsymbol{e}_i\boldsymbol{H}^T && (16.10b) \end{aligned}$$

Knowing $\boldsymbol{r}$, we can compute $\boldsymbol{s}$ [Eq. (16.10a)] and presumably we can compute $\boldsymbol{e}_i$ from

Eq. (16.10b). Unfortunately, knowledge of s does not allow us to solve uniquely for e_i. This is because r can also be expressed in terms of code words other than c_i. Thus,

$$r = c_j \oplus e_j \qquad j \neq i$$

Hence,

$$s = (c_j \oplus e_j)H^T = e_j H^T$$

Because there are 2^k possible code words,

$$s = eH^T$$

is satisfied by 2^k error vectors. To give an example, if a data word $d = \mathbf{100}$ is transmitted by a code word **100101** in Example 16.1, and if a detection error is caused in the third digit, then the received word is **101101**. In this case we have $c = \mathbf{100101}$ and $e = \mathbf{001000}$. But the same word could have been received if $c = \mathbf{101011}$ and $e = \mathbf{000110}$, or if $c = \mathbf{010011}$ and $e = \mathbf{111110}$, and so on. Thus, there are eight possible error vectors (2^k error vectors) that satisfy Eq. (16.10b). Which vector shall we choose? For this, we must define our decision criterion. One reasonable criterion is the maximum-likelihood rule where, if we receive r, then we decide in favor of that c for which r is most likely to be received. In other words, we decide "c_i transmitted" if

$$P(r|c_i) > P(r|c_k) \qquad \text{all } k \neq i$$

For a BSC, this rule gives a very simple answer. Suppose the Hamming distance between r and c_i is d, that is, the channel noise causes errors in d digits. Then if P_e is the digit error probability of a BSC,

$$P(r|c_i) = P_e^d(1 - P_e)^{n-d} = (1 - P_e)^n \left(\frac{P_e}{1 - P_e}\right)^d$$

If $P_e < 0.5$, then $P(r|c_i)$ is a monotonically decreasing function of d because $P_e/(1-P_e) < 1$. Hence, to maximize $P(r|c_i)$, we must choose that c_i which is closest to r; that is, we must choose the error vector with the smallest number of **1**'s. A vector with the smallest number of **1**'s is called the **minimum-weight vector**.

EXAMPLE 16.2 A (6, 3) code is generated according to the generating matrix in Example 16.1. The receiver receives $r = \mathbf{100011}$. Determine the corresponding data word if the channel is a BSC and the maximum-likelihood decision is used.

We have

$$\begin{aligned} s &= rH^T \\ &= [\mathbf{1} \;\; \mathbf{0} \;\; \mathbf{0} \;\; \mathbf{0} \;\; \mathbf{1} \;\; \mathbf{1}] \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= [\mathbf{1} \;\; \mathbf{1} \;\; \mathbf{0}] \end{aligned}$$

Because for modulo-2 operation, subtraction is the same as addition, the correct transmitted code word $\boldsymbol{c}$ is given by

$$\boldsymbol{c} = \boldsymbol{r} \oplus \boldsymbol{e}$$

where $\boldsymbol{e}$ satisfies

$$\begin{aligned} s &= [\mathbf{1} \quad \mathbf{1} \quad \mathbf{0}] = \boldsymbol{e}H^T \\ &= [e_1 \quad e_2 \quad e_3 \quad e_4 \quad e_5 \quad e_6] \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

We see that $\boldsymbol{e} = \mathbf{001000}$ satisfies this equation. But so does $\boldsymbol{e} = \mathbf{000110}$, or **010101**, or **011011**, or **111110**, or **110000**, or **101101**, or **100011**. The suitable choice, the minimum-weight $\boldsymbol{e}$, is **001000**. Hence,

$$\boldsymbol{c} = \mathbf{100011} \oplus \mathbf{001000} = \mathbf{101011}$$

The decoding procedure just described is quite disorganized. A systematic procedure would be to consider all possible syndromes and for each syndrome associate a minimum-weight error vector. For instance, the single-error correcting code in Example 16.1 has a syndrome with three digits. Hence, there are eight possible syndromes. We prepare a table of minimum-weight error vectors corresponding to each syndrome (see Table 16.3). This table can be prepared by considering all possible minimum-weight error vectors and computing s for each of them using Eq. (16.10b). The first minimum-weight error vector **000000** is a trivial case that has the syndrome **000**. Next, we consider all possible unit weight error vectors. There are six such vectors: **100000, 010000, 001000, 000100, 000010, 000001**. Syndromes for these can readily be calculated from Eq. (16.10b) and tabulated (Table 16.3). This still leaves one syndrome, **111**, that is not matched with some error vector. Since all unit-weight error vectors are exhausted, we must look for error vectors of weight 2.

We find that for the first seven syndromes (Table 16.3), there is a unique minimum-weight vector $\boldsymbol{e}$. But for $s = \mathbf{111}$, the error vector $\boldsymbol{e}$ has a minimum weight of 2, and it is not unique. For example, $\boldsymbol{e} = \mathbf{100010}$ or **010100** or **001001** all have $s = \mathbf{111}$, and all three $\boldsymbol{e}$'s are minimum weight (weight 2). In such a case, we can pick any one of these $\boldsymbol{e}$'s as a **correctable** error pattern. In Table 16.3, we have picked $\boldsymbol{e} = \mathbf{100010}$ as the double-error correctable pattern. This means the present code can correct all six single-error patterns and one double-error pattern (**100010**). For instance, if $\boldsymbol{c} = \mathbf{101011}$ is transmitted and the channel noise causes the double error **100010**, the received vector $\boldsymbol{r} = \mathbf{001001}$, and

$$s = \boldsymbol{r}\boldsymbol{H}^T = [\mathbf{1}\,\mathbf{1}\,\mathbf{1}]$$

From Table 16.3 we see that corresponding to $s = \mathbf{111}$ is $\boldsymbol{e} = \mathbf{100010}$, and we immediately decide $\boldsymbol{c} = \boldsymbol{r} \oplus \boldsymbol{e} = \mathbf{101011}$. Note, however, that this code will not correct double-error patterns other than **100010**. Thus, this code not only corrects all single errors but one double-error pattern as well. This extra bonus of one double-error correction occurs because n and

Table 16.3
Decoding table for code in Table 16.2

e	s
000000	000
100000	101
010000	011
001000	110
000100	100
000010	010
000001	001
100010	111

k oversatisfy the Hamming bound [Eq. (16.2b)]. In case n and k were to satisfy the bound exactly, we would have only single-error correction ability. This is the case for the (7, 4) code, which can correct all single-error patterns only.

Thus for systematic decoding, we prepare a table of all correctable error patterns and the corresponding syndromes. For decoding, we need only calculate $s = \boldsymbol{r}\boldsymbol{H}^T$ and, from the decoding table, find the corresponding $\boldsymbol{e}$. The decision is $\boldsymbol{c} = \boldsymbol{r} \oplus \boldsymbol{e}$.

Because s has $m = n - k$ digits, there is a total of 2^{n-k} syndromes, each of $n - k$ digits. There is the same number of correctable error vectors, each of n digits. Hence, for the purpose of decoding, we need a storage of $(2n - k)2^{n-k} = (2n - k)2^m$. This storage requirement grows exponentially with m, the number of parity-check digits, and can be enormous, even for moderately complex codes.

Because the maximum-likelihood decision is the same as choosing the code word closest to the received word, we could just as well compare the received word with each of the 2^k possible code words of n digits each. This involves a storage of $n2^k$, which can be much larger than the $(2n - k)2^m$ storage required earlier.

It is still not clear how to choose coefficients of the generator or parity-check matrix. Unfortunately, there is no systematic way to do this, except for the case of single-error correcting codes, also known as *Hamming codes*. Let us consider a single-error correcting (7, 4) code. This code satisfies the Hamming bound exactly, and we shall see that a proper code can be constructed. In this case $m = 3$, and there are seven nonzero syndromes, and because $n = 7$, there are exactly seven single-error patterns. Hence, we can correct all single-error patterns and no more. Consider the single-error pattern $\boldsymbol{e} = \mathbf{1000000}$. Because

$$s = \boldsymbol{e}\boldsymbol{H}^T$$

$\boldsymbol{e}\boldsymbol{H}^T$ will be simply the first row of $\boldsymbol{H}^T$. Similarly, for $\boldsymbol{e} = \mathbf{0100000}$, $s = \boldsymbol{e}H^T$ will be the second row of $\boldsymbol{H}^T$, and so on. Now for unique decodability, we require that all seven syndromes corresponding to the seven single-error patterns be distinct. Conversely, if all the seven syndromes are distinct, we can decode all the single-error patterns. This means that the only requirement on $\boldsymbol{H}^T$ is that all seven of its rows be distinct and nonzero. Note that $\boldsymbol{H}^T$ is an $(n \times n - k)$ matrix (i.e., 7×3 in this case). Because there exist seven nonzero patterns of three digits, it is possible to find seven nonzero rows of three digits each. There are many ways in which these rows can be ordered. But we must remember that the three bottom rows must form identity matrix $\boldsymbol{I}_m$ [see Eq. (16.9b)].

One possible form of $\boldsymbol{H}^T$ is

$$H^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{P} \\ \boldsymbol{I}_m \end{bmatrix}$$

The corresponding generator matrix $\boldsymbol{G}$ is

$$\boldsymbol{G} = [\boldsymbol{I}_k \quad \boldsymbol{P}] = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

Thus when $\boldsymbol{d} = \mathbf{1011}$, the corresponding code word $\boldsymbol{c} = \mathbf{1011001}$, and so forth.

A general (n, k) code has m-dimensional syndrome vectors $(m = n - k)$. Hence, there are $2^m - 1$ distinct nonzero syndrome vectors that can correct $2^m - 1$ single-error patterns. Because in an (n, k) code there are exactly n single-error patterns, all these patterns can be corrected if

$$2^m - 1 \geq n$$

or

$$2^{n-k} \geq n + 1$$

This is precisely the condition in Eq. (16.3) for $t = 1$. Thus, for any (n, k) satisfying this condition, it is possible to construct a single-error correcting code by the procedure discussed.

More discussion on block coding can be found in Peterson and Weldon[1] as well as in Lin and Costello.[2]

16.3 CYCLIC CODES[1,2]

Cyclic codes are a subclass of linear block codes. As seen before, a procedure for selecting a generator matrix is relatively easy for single-error correcting codes. This procedure, however, cannot carry us very far in constructing higher order error correcting codes. Cyclic codes have a fair amount of mathematical structure that permits the design of higher order correcting codes. Second, for cyclic codes, encoding and syndrome calculations can be easily implemented using simple shift registers.

Cyclic codes are such that code words are simple lateral shifts of one another. For example, if $\boldsymbol{c} = (c_1, c_2, \ldots, c_{n-1}, c_n)$ is a code word, then so are $(c_2, c_3, \ldots, c_n, c_1)$ and $(c_3, c_4, \ldots, c_n, c_1, c_2)$, and so on. We shall use the following notation. If

$$\boldsymbol{c} = (c_1, c_2, \ldots, c_n) \tag{16.11a}$$

is a code vector of a code C, then $\boldsymbol{c}^{(i)}$ denotes $\boldsymbol{c}$ shifted cyclically i places to the left, that is,

$$\boldsymbol{c}^{(i)} = (c_{i+1}, c_{i+2}, \ldots, c_n, c_1, c_2, \ldots, c_i) \tag{16.11b}$$

Cyclic codes can be described in a polynomial form. This property is extremely useful in the analysis and implementation of these codes. The code vector $\boldsymbol{c}$ in Eq. (16.11a) can be expressed as the $(n-1)$-degree polynomial

$$c(x) = c_1x^{n-1} + c_2x^{n-2} + \cdots + c_n \tag{16.12a}$$

The coefficients of the polynomial are either **0** or **1**, and they obey the following properties:

$$\begin{array}{ll} \mathbf{0}+\mathbf{0}=\mathbf{0} & \mathbf{0}\times\mathbf{0}=\mathbf{0} \\ \mathbf{0}+\mathbf{1}=\mathbf{1}+\mathbf{0}=\mathbf{1} & \mathbf{0}\times\mathbf{1}=\mathbf{1}\times\mathbf{0}=\mathbf{0} \\ \mathbf{1}+\mathbf{1}=\mathbf{0} & \mathbf{1}\times\mathbf{1}=\mathbf{1} \end{array}$$

The code polynomial $c^{(i)}(x)$ for the code vector $\boldsymbol{c}^{(i)}$ in Eq. (16.11b) is

$$c^{(i)}(x) = c_{i+1}x^{n-1} + c_{i+2}x^{n-2} + \cdots + c_nx^i + c_1x^{i-1} + \cdots + c_i \tag{16.12b}$$

One of the interesting properties of the code polynomials is that when $x^ic(x)$ is divided by x^n+1, the remainder is $c^{(i)}(x)$. We can verify this property as follows:

$$xc(x) = c_1x^n + c_2x^{n-1} + \cdots + c_nx$$

$$\begin{array}{rl} & c_1 \\ x^n+1\big) & \overline{c_1x^n + c_2x^{n-1} + \cdots + c_nx} \\ & c_1x^n + c_1 \\ \hline & c_2x^{n-1} + c_3x^{n-2} + \cdots + c_nx + c_1 \quad \leftarrow \text{remainder} \end{array}$$

The remainder is clearly $c^{(1)}(x)$. In deriving this result, we have used the fact that subtraction amounts to summation when modulo-2 operations are involved. Continuing in this fashion, we can show that the remainder of $x^ic(x)$ divided by x^n+1 is $c^{(i)}(x)$.

We shall now prove an important theorem in cyclic codes. It says that a cyclic code polynomial $c(x)$ can be generated by the data polynomial $d(x)$ of degree $k-1$ and a generator polynomial $g(x)$ of degree $n-k$ as

$$c(x) = d(x)\,g(x) \tag{16.13}$$

where the generator polynomial $g(x)$ is an $(n-k)$th-order factor of (x^n+1).

For a data vector $(d_1, d_2, \ldots, d_k)$, the data polynomial is

$$d(x) = d_1x^{k-1} + d_2x^{k-2} + \cdots + d_k \tag{16.14}$$

Proof: Consider a polynomial

$$\begin{aligned} c(x) &= d(x)\,g(x) \\ &= d_1x^{k-1}g(x) + d_2x^{k-2}g(x) + \cdots + d_kg(x) \end{aligned} \tag{16.15}$$

This is a polynomial of degree $n-1$ or less. There are a total of 2^k such polynomials corresponding to 2^k data vectors. Thus, we have a linear (n, k) code generated by Eq. (16.13). To prove that this code is cyclic, let

$$c(x) = c_1x^{n-1} + c_2x^{n-2} + \cdots + c_n$$

be a code polynomial in this code [Eq. (16.15)]. Then,

$$\begin{aligned} xc(x) &= c_1x^n + c_2x^{n-1} + \cdots + c_nx \\ &= c_1(x^n + 1) + (c_2x^{n-1} + c_3x^{n-2} + \cdots + c_nx + c_1) \\ &= c_1(x^n + 1) + c^{(1)}(x) \end{aligned}$$

Because $xc(x)$ is $xd(x)g(x)$, and $g(x)$ is a factor of $x^n + 1$, $c^{(1)}(x)$ must also be a multiple of $g(x)$ and can also be expressed as $d(x)g(x)$ for some data vector $\mathbf{d}$. Therefore, $c^{(1)}(x)$ is also a code polynomial. Continuing this way, we see that $c^{(2)}(x), c^{(3)}(x), \ldots$ are all code polynomials generated by Eq. (16.15). Thus, the linear (n, k) code generated by $d(x)g(x)$ is indeed cyclic.

EXAMPLE 16.3 Find a generator polynomial $g(x)$ for a $(7, 4)$ cyclic code, and find code vectors for the following data vectors: **1010**, **1111**, **0001**, and **1000**.

In this case $n = 7$ and $n - k = 3$, and

$$x^7 + 1 = (x + 1)(x^3 + x + 1)(x^3 + x^2 + 1)$$

For a $(7, 4)$ code, the generator polynomial must be of the order $n - k = 3$. In this case, there are two possible choices for $g(x)$: $x^3 + x + 1$ or $x^3 + x^2 + 1$. Let us choose the latter, that is,

$$g(x) = x^3 + x^2 + 1$$

as a possible generator polynomial. For $\mathbf{d} = [\mathbf{1} \quad \mathbf{0} \quad \mathbf{1} \quad \mathbf{0}]$,

$$d(x) = x^3 + x$$

and the code polynomial is

$$\begin{aligned} c(x) &= d(x)g(x) \\ &= (x^3 + x)(x^3 + x^2 + 1) \\ &= x^6 + x^5 + x^4 + x \end{aligned}$$

Hence,

$$c = \mathbf{1110010}$$

Similarly, code words for other data words can be found (see Table 16.4).

Table 16.4

d	***e***
1010	**1110010**
1111	**1001011**
0001	**0001101**
1000	**1101000**

Note the structure of the code words. The first k digits are not necessarily the data digits. Hence, this is not a systematic code.

In a systematic code, the first k digits are data digits, and the last $m = n - k$ digits are the parity-check digits. Systematic codes are a special case of general codes. Our discussion thus far applies to general cyclic codes, of which systematic cyclic codes are a special case. We shall now develop a method of generating systematic cyclic codes.

Systematic Cyclic Codes

We shall show that for a systematic code, the code word polynomial $c(x)$ corresponding to the data polynomial $d(x)$ is given by

$$c(x) = x^{n-k}d(x) + \rho(x) \tag{16.16a}$$

where $\rho(x)$ is the remainder from dividing $x^{n-k}d(x)$ by $g(x)$,

$$\rho(x) = \text{Rem}\,\frac{x^{n-k}d(x)}{g(x)} \tag{16.16b}$$

To prove this we observe that

$$\frac{x^{n-k}d(x)}{g(x)} = q(x) + \frac{\rho(x)}{g(x)} \tag{16.17a}$$

where $q(x)$ is of degree $k - 1$ or less. We add $\rho(x)/g(x)$ to both sides of Eq. (16.17a), and because $f(x) + f(x) = 0$ under modulo-2 operation, we have

$$\frac{x^{n-k}d(x) + \rho(x)}{g(x)} = q(x) \tag{16.17b}$$

or

$$q(x)g(x) = x^{n-k}d(x) + \rho(x) \tag{16.17c}$$

Because $q(x)$ is on the order of $k - 1$ or less, $q(x)g(x)$ is a code polynomial. As $x^{n-k}d(x)$ represents $d(x)$ shifted to the left by $n-k$ digits, the first k digits of this code word are precisely $\boldsymbol{d}$, and the last $n-k$ digits corresponding to $\rho(x)$ must be parity-check digits. This will become clear by considering a specific example.

EXAMPLE 16.4 Construct a systematic (7, 4) cyclic code using a generator polynomial (see Example 16.3).

We use

$$g(x) = x^3 + x^2 + 1$$

Consider a data vector $\boldsymbol{d} = \mathbf{1010}$,

$$d(x) = x^3 + x$$

and

$$x^{n-k}d(x) = x^6 + x^4$$

Hence,

$$
\begin{array}{r l}
 & \underline{x^3 + x^2 + 1} \qquad \leftarrow q(x) \\
x^3 + x^2 + 1 \,\big) & x^6 + x^4 \\
 & \underline{x^6 + x^5 + x^3} \\
 & \quad x^5 + x^4 + x^3 \\
 & \quad \underline{x^5 + x^4 + x^2} \\
 & \qquad x^3 + x^2 \\
 & \qquad \underline{x^3 + x^2 + 1} \\
 & \qquad\qquad\quad 1 \qquad \leftarrow \rho(x)
\end{array}
$$

Hence, from Eq. (16.16a),

$$
\begin{aligned}
c(x) &= x^3 d(x) + \rho(x) \\
&= x^3(x^3 + x) + 1 \\
&= x^6 + x^4 + 1
\end{aligned}
$$

and

$$\boldsymbol{c} = \mathbf{1010001}$$

We could also have found the code word $\boldsymbol{c}$ directly by using Eq. (16.17c). Thus, $c(x) = q(x)g(x) = (x^3+x^2+1)(x^3+x^2+1) = x^6+x^4+1$. We construct the entire code table in this manner (Table 16.5). This is quite a laborious procedure. There is, however, a shortcut, using the code generating matrix $\boldsymbol{G}$. Using the earlier procedure, we compute the code words corresponding to the data words **1000**, **0100**, **0010**, **0001**. These are **1000110**, **0100011**, **0010111**, **0001101**. Now recognize that these four code words are the four rows of $\boldsymbol{G}$. This is because $\boldsymbol{c} = \boldsymbol{d}\boldsymbol{G}$, and when $\boldsymbol{d} = \mathbf{1000}$, $\boldsymbol{d}\boldsymbol{G}$ is the first row of $\boldsymbol{G}$, and so on. Hence,

$$
\boldsymbol{G} = \begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 1
\end{bmatrix}
$$

Now, we can construct the rest of the code table using $\boldsymbol{c} = \boldsymbol{d}\boldsymbol{G}$. This is an efficient method because it allows us to construct the entire code table from the knowledge of only n code words.

Table 16.5 shows the complete code. Note that $d_{\min}$, the minimum distance between two code words, is 3. Hence, this is a single-error correcting code, and 14 of these code words can be obtained by successive cyclic shifts of the two code words **1110010** and **1101000**. The remaining two code words **1111111** and **0000000** remain unchanged under cyclic shift.

Cyclic codes can also be described by a generator matrix $\boldsymbol{G}$ (see Probs. 16.3-6 and 16.3-7). It can be shown that Hamming codes are cyclic codes.

Table 16.5

d	c
1111	1111111
1110	1110010
1101	1101000
1100	1100101
1011	1011100
1010	1010001
1001	1001011
1000	1000110
0111	0111001
0110	0110100
0101	0101110
0100	0100011
0011	0011010
0010	0010111
0001	0001101
0000	0000000

Cyclic Code Generation

One of the advantages of cyclic codes is that coding and decoding can be implemented using such simple elements as shift registers and modulo-2 adders. A systematic code generation is described in Eqs. (16.16). This involves division of $x^{n-k}d(x)$ by $g(x)$ and can be implemented by a dividing circuit, which is a shift register with feedback connections according to the generator polynomial* $g(x) = x^{n-k} + g_1 x^{n-k-1} + \cdots + g_{n-k-1}x + 1$. The gain g_k's are either 0 or 1. An encoding circuit with $n-k$ shift registers is shown in Fig. 16.2. An understanding of this dividing circuit requires some background in linear sequential networks. An explanation of its functioning can be found in Peterson and Weldon.[1] The k data digits are shifted in one at a time at the input with the switch s held at position p_1. The symbol D represents a one-digit delay. As the data digits move through the encoder, they are also shifted out onto the output line, because the first k digits of the code word are the data digits themselves. As soon as the last (or kth) data digit clears the last [or $(n-k)$th] register, all the registers contain the parity-check digits. The switch s is now thrown to position p_2, and the $n-k$ parity-check digits are shifted out one at a time onto the line.

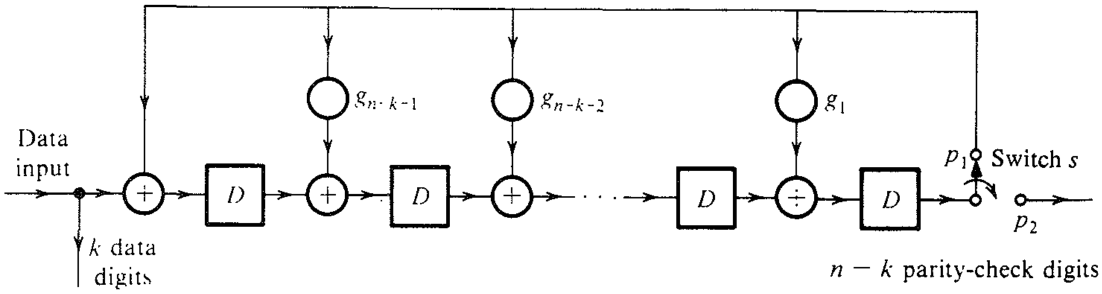


Figure 16.2 Encoder for systematic cyclic code.

---

* It can be shown that for cyclic codes, the generator polynomial must be of this form.

Decoding

Every valid code polynomial $c(x)$ is a multiple of $g(x)$. If an error occurs during the transmission, the received word polynomial $r(x)$ will not be a multiple* of $g(x)$. Thus,

$$\frac{r(x)}{g(x)} = m_1(x) + \frac{s(x)}{g(x)} \tag{16.18}$$

and

$$s(x) = \text{Rem}\,\frac{r(x)}{g(x)} \tag{16.19}$$

where the syndrome polynomial $s(x)$ has a degree $n - k - 1$ or less.

If $e(x)$ is the error polynomial, then

$$r(x) = c(x) + e(x)$$

Remembering that $c(x)$ is a multiple of $g(x)$,

$$\begin{aligned} s(x) &= \text{Rem}\,\frac{r(x)}{g(x)} \\ &= \text{Rem}\,\frac{c(x) + e(x)}{g(x)} \\ &= \text{Rem}\,\frac{e(x)}{g(x)} \end{aligned} \tag{16.20}$$

Again, as before, a received word $\boldsymbol{r}$ could result from any one of the 2^k code words and a suitable error. For example, for the code in Table 16.5, if $\boldsymbol{r}$ = **0110010**, this could mean $\boldsymbol{c}$ = **1110010** and $\boldsymbol{e}$ = **1000000**, or $\boldsymbol{c}$ = **1101000** and $\boldsymbol{e}$ = **1011010**, or 14 more such combinations. As seen earlier, the most likely error pattern is the one with the minimum weight (or minimum number of 1's). Hence, here $\boldsymbol{c}$ = **1110010** and $\boldsymbol{e}$ = **1000000** is the correct decision.

It is convenient to prepare a decoding table, that is, to list the syndromes for all correctable errors. For any $\boldsymbol{r}$, we compute the syndrome from Eq. (16.19), and from the table we find the corresponding correctable error $\boldsymbol{e}$. Then we determine $\boldsymbol{c} = \boldsymbol{r} \oplus \boldsymbol{e}$. Note that computation of $s(x)$ [see Eq. (16.18)] involves exactly the same operation as that required to compute $\rho(x)$ in coding [see Eq. (16.17a)]. Hence, the circuit in Fig. 16.2 can also be used to compute $s(x)$.

EXAMPLE 16.5 Construct the decoding table for the single-error correcting (7, 4) code in Table 16.5. Determine the data vectors transmitted for the following received vectors $\boldsymbol{r}$: (a) **1101101**; (b) **0101000**; (c) **0001100**.

The first step is to construct the decoding table. Because $n - k - 1 = 2$, the syndrome polynomial is of the second order, and there are seven possible nonzero syndromes. There are also seven possible correctable single-error patterns because $n = 7$. Using Eq. (16.20), we compute the syndrome for each of the seven correctable error patterns. For example, for $\boldsymbol{e}$ = **1000000**, $e(x) = x^6$. Because $g(x) = x^3 + x^2 + 1$ for this code (see Example 16.4), we have

---

* This assumes that the number of errors in **r** is correctable.

$$\begin{array}{r r}
 & x^3 + x^2 + x \\
x^3 + x^2 + 1 \big) & \overline{x^6 \qquad\qquad\qquad\qquad\qquad} \\
 & \underline{x^6 + x^5 + x^3 \qquad\qquad\qquad} \\
 & x^5 + x^3 \qquad\qquad \\
 & \underline{x^5 + x^4 + x^2 \qquad} \\
 & x^4 + x^3 + x^2 \\
 & \underline{x^4 + x^3 + x} \\
 & x^2 + x \quad \leftarrow s(x)
\end{array}$$

Hence,

$$\mathbf{s} = \mathbf{110}$$

In a similar way, we compute the syndromes for the remaining error patterns (see Table 16.6).

Table 16.6

e	***s***
1000000	**110**
0100000	**011**
0010000	**111**
0001000	**101**
0000100	**100**
0000010	**010**
0000001	**001**

When the received word $\mathbf{r}$ is **1101101**,

$$r(x) = x^6 + x^5 + x^3 + x^2 + 1$$

We now compute $s(x)$ according to Eq. (16.19):

$$\begin{array}{r l}
 & x^3 \\
x^3 + x^2 + 1 \big) & \overline{x^6 + x^5 + x^3 + x^2 + 1} \\
 & \underline{x^6 + x^5 + x^3 \qquad\qquad\quad} \\
 & \qquad\qquad\qquad\quad x^2 + 1
\end{array}$$

Hence, $\mathbf{s} = \mathbf{101}$. From Table 16.6, this gives $\mathbf{e} = \mathbf{0001000}$, and

$$\mathbf{c} = \mathbf{r} \oplus \mathbf{e} = \mathbf{1101101} \oplus \mathbf{0001000} = \mathbf{1100101}$$

Hence, from Table 16.5 we have

$$\mathbf{d} = \mathbf{1100}$$

In a similar way, we determine for $\mathbf{r} = \mathbf{0101000}$, $\mathbf{s} = \mathbf{110}$ and $\mathbf{e} = \mathbf{1000000}$; hence $\mathbf{c} = \mathbf{r} \oplus \mathbf{e} = \mathbf{1101000}$, and $\mathbf{d} = \mathbf{1101}$. For $\mathbf{r} = \mathbf{0001100}$, $\mathbf{s} = \mathbf{001}$ and $\mathbf{e} = \mathbf{0000001}$; hence $\mathbf{c} = \mathbf{r} \oplus \mathbf{e} = \mathbf{0001101}$, and $\mathbf{d} = \mathbf{0001}$.

Bose-Chaudhuri-Hocquenghen (BCH) Codes

The BCH codes are perhaps the most powerful of the random-error correcting cyclic codes. Moreover, their decoding procedure can be implemented simply. Hamming code is a special case of BCH codes. These codes are described as follows: For any positive integers m' and t ($t < 2^{m'-1}$), there exists a t-error correcting (n, k) code with $n = 2^{m'} - 1$ and $n - k \leq m't$. The minimum distance $d_{\min}$ between code words is related by the inequality $2t + 1 \leq d_{min} \leq 2t + 2$.

The detailed treatment of BCH codes requires extensive use of modern algebra and is beyond our scope. For further discussion of BCH codes, the reader is referred to Lin and Costello[2] or Peterson and Weldon.[1]

16.4 BURST-ERROR DETECTING AND CORRECTING CODES

Thus far we have considered detecting or correcting errors that occur independently, or randomly, in digit positions. On some channels, disturbances can wipe out an entire block of digits. For instance, a stroke of lightning or a human-made electrical disturbance can affect several adjacent transmitted digits. On magnetic storage systems, magnetic tape defects usually affect more than one digit. Burst errors are those errors that wipe out some or all of a sequential set of digits. In general, random-error correcting codes are not efficient for correcting burst errors. Hence, special **burst-error correcting codes** are used for this purpose.

A burst of length b is defined as a sequence of digits in which the first digit and the bth digit are in error, with the $b - 2$ digits in between either in error or received correctly. For example, an error vector $\boldsymbol{e} = \mathbf{0010010100}$ has a burst length of 6.

It can be shown that for detecting all burst errors of length b or less with a linear block code of length n, b parity-check bits are necessary and sufficient.[1] We shall prove the sufficiency part of this theorem by constructing a code of length n with b parity-check digits that will detect a burst of length b.

To construct such a code, let us group k data digits into segments of b digits in length (Fig. 16.3). To this we add a last segment of b parity-check digits, which are determined as follows. The modulo-2 sum of the ith digit in each segment (including the parity-check segment) must be zero. For example, the first digits in the five data segments are **1**, **0**, **1**, **1**, and **1**. Hence, we must have **0** as the first parity-check digit in order to obtain a modulo-2 sum zero. We continue in this way with the second digit, the third digit, and so on, to the bth digit. Because parity-check digits are a linear combination of data digits, this is a linear block code. Moreover, it is a systematic code.

It is easy to see that if a digit sequence of length b or less is in error, parity will be violated and the error will be detected (but not corrected), and the receiver can request retransmission of the digits lost. One of the interesting properties of this code is that b, the number of parity-check digits, is independent of k (or n), which makes it a very useful code for such systems as

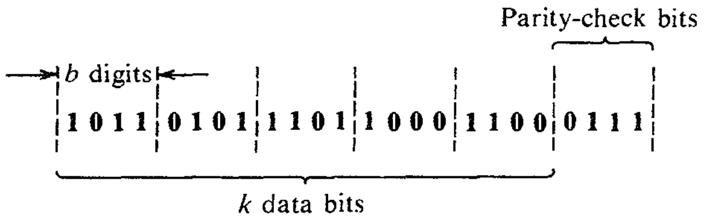


Figure 16.3 Burst-error detection.

packet switching, where the data digits may vary from packet to packet. It can be shown that a linear code with b parity bits detects not only all bursts of length b or less, but also a high percentage of longer bursts.[1]

If we are interested in correcting rather than detecting burst errors, we require twice as many parity-check digits. A theorem says: In order to correct all burst errors of length b or less, a linear block code must have at least $2b$ parity-check digits.[1]

16.5 INTERLACED CODES FOR BURST- AND RANDOM-ERROR CORRECTION

In general, random-error correcting codes are not efficient for burst-error correcting, and burst-error correcting codes are not efficient for random-error correcting. Unfortunately, in most practical systems, we have errors of both kinds. Out of the several methods proposed to simultaneously correct random and burst errors, the interlaced code is the most effective.

For an (n, k) code, if we interlace λ code words, we have what is known as a $(\lambda n, \lambda k)$ **interlaced code**. Instead of transmitting code words one by one, we group λ code words and interlace them. Consider, for example, the case of $\lambda = 3$ and a two-error correcting $(15, 8)$ code. Each code word has 15 digits. We group code words to be transmitted in groups of three. Suppose the first three code words to be transmitted are $\boldsymbol{x}$ $(x_1, x_2, \ldots, x_{15})$, $\boldsymbol{y}$ $(y_1, y_2, \ldots, y_{15})$, and $\boldsymbol{z}$ $(z_1, z_2, \ldots, z_{15})$, respectively. Then instead of transmitting $\boldsymbol{xyz}$ in sequence as $x_1, x_2, \ldots, x_{15}, y_1, y_2, \ldots, y_{15}, z_1, z_2, \ldots, z_{15}$, we transmit $x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3, \ldots, x_{15}, y_{15}, z_{15}$. This can be explained graphically by Fig. 16.4, where λ code words (three in this case) are arranged in rows. In usual transmission, we transmit one row after another. In the interlaced case, we transmit columns (of λ elements) in sequence. When all the 15 (n) columns are transmitted, we repeat the procedure for the next λ code words to be transmitted.

To explain the error correcting capabilities of this code, we observe that the decoder will first remove the interlacing and regroup the received digits as $x_1, x_2, \ldots, x_{15}, y_1, y_2, \ldots, y_{15}, z_1, z_2, \ldots, z_{15}$. Suppose the shaded digits in Fig. 16.4 were in error. Because the code is a two-error correcting code, two or less errors in each row will be corrected. Hence, all the errors in Fig. 16.4 are correctable. We see that there are two random, or independent, errors and one burst of length 4 in all the 45 digits transmitted. In general, if the original (n, k) code is t-error correcting, the interlaced code can correct any combination of t bursts of length λ or less.

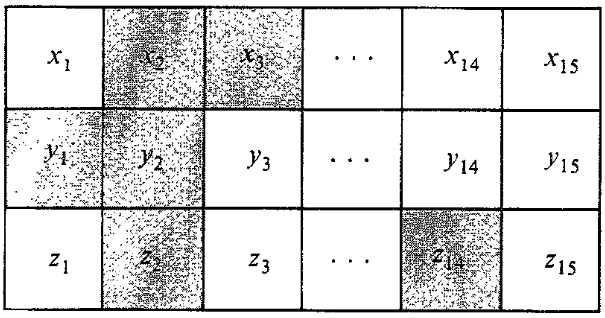


Figure 16.4 Random- and burst-error correction.

16.6 CONVOLUTIONAL CODES

Convolutional (or recurrent) codes, first introduced in 1955 by Elias,[3] differ from block codes as follows. In a block code, the block of n code digits generated by the encoder in any particular time unit depends only on the block of k input data digits within that time unit. In a convolutional code, on the other hand, the block of n code digits generated by the encoder in a particular time unit depends not only on the block of k message digits within that time unit but also on the block of data digits within a previous span of $N-1$ time units ($N > 1$). For convolutional codes, k and n are usually small. Convolutional codes can be devised for correcting random errors, burst errors, or both. Encoding is easily implemented by shift registers. As a class, convolutional codes invariably outperform block codes of the same order of complexity.

A convolutional coder with constraint length N consists of an N-stage shift register and v modulo-2 adders. Figure 16.5 shows such a coder for the case of $N = 3$ and $v = 2$. The message digits are applied at the input of the shift register. The coded digit stream is obtained at the commutator output. The commutator samples the v moldulo-2 adders in sequence, once during each input-bit interval. We shall explain this operation with reference to the input digits **11010**. Initially, all the stages of the register are clear; that is, they are in a **0** state. When the first data digit **1** enters the register, the stage s_1 shows **1** and all the other stages (s_2 and s_3) are unchanged; that is, they are in a **0** state. The two modulo-2 adders show $v_1 = \mathbf{1}$ and $v_2 = \mathbf{1}$. The commutator samples this output. Hence, the coder output is **11**. When the second message bit **1** enters the register, it enters the stage s_1, and the previous **1** in s_1 is shifted to s_2. Hence, s_1 and s_2 both show **1**, and s_3 is still unchanged; that is, it is in a **0** state. The modulo-2 adders now show $v_1 = \mathbf{0}$ and $v_2 = \mathbf{1}$. Hence, the decoder output is **01**. In the same way, when the third message digit **0** enters the register, we have $s_1 = \mathbf{0}$, $s_2 = \mathbf{1}$, and $s_3 = \mathbf{1}$, and the decoder output is **01**. Observe that each data digit influences N groups of v digits in the output (in this case three groups of two digits). The process continues until the last data digit enters the stage s_1.* We cannot stop here, however. We add $N-1$ number of **0**'s to the input stream (dummy or augmented data) to make sure that the last data digit (**0** in this case) proceeds all the way through the shift register in order to influence the N groups of v digits. Hence, when the input digits are **11010**, we actually apply **1101000** (the digits augmented by $N-1$ zeros) to the

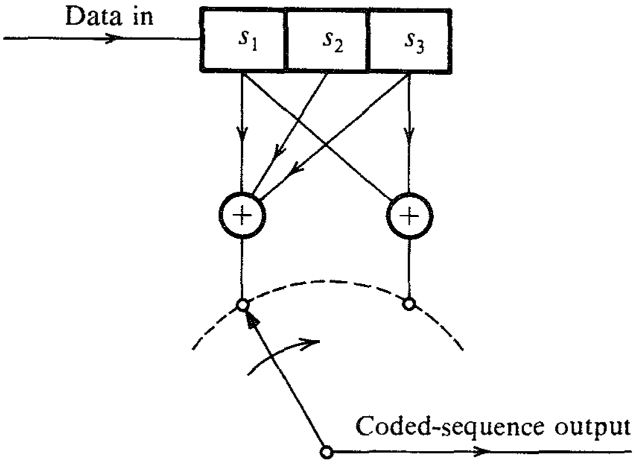


Figure 16.5 Convolutional coder.

---

* For a systematic code, one of the output digits must be the data digit itself.

input of the shift register. It can be seen that when the last digit of the augmented message stream enters s_1, the last digit of the message stream has passed through all the N stages of the register. The reader can verify that the coder output is given by **11 01 01 00 10 11 00**. Thus, there are in all $n = (N + k - 1)v$ digits in the coded output for every k data digits. In practice, $k \gg N$, and, hence, there are approximately kv coded output digits for every k data digits, giving an efficiency $\eta \simeq 1/v$.*

It can be seen that unlike the block coder, the convolutional coder operates on a continuous basis, and each data digit influences N groups of v digits in the output.

Code Tree

Coding and decoding is considerably facilitated by what is known as the **code tree**, which shows the coded output for any possible sequence of data digits. The code tree for the coder in Fig. 16.5 with $k = 5$ is shown in Fig. 16.6. When the first digit is **0**, the coder output is **00**, and when it is **1**, the output is **11**. This is shown by the two tree branches that start at the initial node. The upper branch represents **0**, and the lower branch represents **1**. This convention will be followed throughout. At the terminal node of each of the two branches, we follow a similar procedure, corresponding to the second data digit. Hence, two branches initiate from each node, the upper one for **0** and the lower one for **1**. This continues until the kth data digit. From there on, all the input digits are **0** (augmented digit), and we have only one branch until the end. Hence, in all there are 32 (or 2^k) outputs corresponding to 2^k possible data vectors. The coded output for input **11010** can be easily read from this tree (the path shown dashed in Fig. 16.6).

Figure 16.6 shows that the code tree becomes repetitive after the third branch. This can be seen from the fact that the two blocks enclosed inside the dashed lines are identical. It means that the output from the fourth input digit is the same whether the first digit was **1** or **0**. This is not surprising in view of the fact that when the fourth input digit enters the shift register, the first input digit is shifted out of the register, and it ceases to influence the output digits. In other words, the data vector $\mathbf{1}x_1x_2x_3x_4\cdots$ and the data vector $\mathbf{0}x_1x_2x_3x_4\cdots$ generate the same output after the third group of output digits. It is convenient to label the four third-level nodes (the nodes appearing at the beginning of the third branch) as nodes a, b, c, and d (Fig. 16.6). The repetitive structure begins at the fourth-level nodes and continues at the fifth-level nodes, whose behavior is similar to that of nodes a, b, c, and d at the third level. Hence, we label the fourth- and fifth-level nodes also as either a, b, c, or d. What this means is that at the fifth-level nodes, the first two data digits have become irrelevant; that is, any of the four combinations (**11**, **10**, **01**, or **00**) for the first two data digits will give the same output after the fifth node.

This behavior can be seen from another point of view. When a data bit enters the shift register (in stage s_1), the output bits are determined not only by the data bit in s_1, but by the two previous data bits already in stages s_3 and s_2. There are four possible combinations of the two previous bits (in s_3 and s_2): **00**, **01**, **10**, and **11**. We shall label these four states as a, b, c, and d, respectively, as shown in Fig. 16.7a. Thus, when the previous two bits are **01** ($s_3 = \mathbf{0}$, $s_2 = \mathbf{1}$), the state is b, and so on. The number of states is equal to 2^{N-1}.

A data bit **0** or **1** generates four different outputs, depending on the encoder state. If the data bit is **0**, the encoder output is **00**, **10**, **11**, or **01**, depending on whether the encoder state is

---

* In general, instead of shifting one digit at a time, b digits may be shifted at a time. In this case $\eta \simeq b/v$.

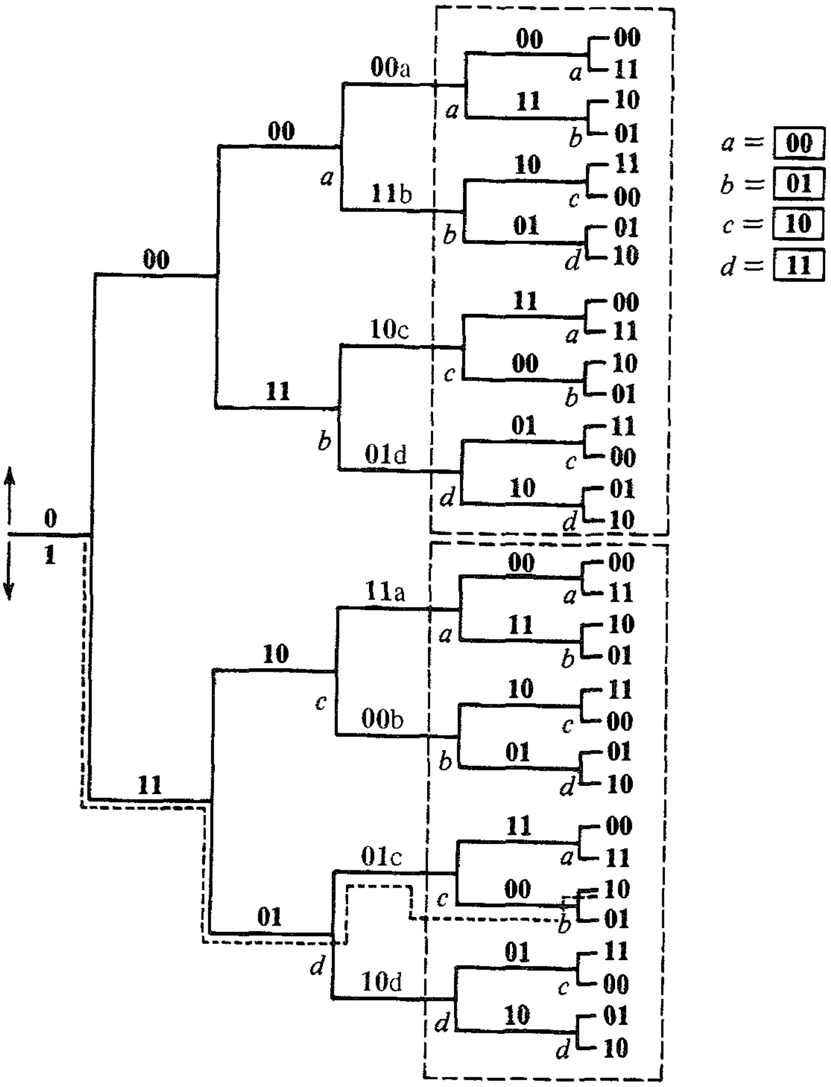


Figure 16.6 Code tree for the coder in Fig. 16.5.

a, *b*, *c*, or *d*. Similarly if the data bit is **1**, the encoder output is **11**, **01**, **00**, or **10**, depending on whether the encoder state is *a*, *b*, *c*, or *d*. This entire behavior can be concisely expressed by the state diagram shown in Fig. 16.7b. This is a four-state directed graph used to uniquely represent the input-output relation of this encoder. We use solid lines when the input bit is **0**, and dashed lines when the input bit is **1**. For instance, when the encoder is in state *a*, and we input **1**, the encoder output is **11** (dashed line). The encoder now goes to state *b* for the next data bit because at this point the previous two bits become $s_3 = \mathbf{0}$ and $s_2 = \mathbf{1}$. Similarly, when the encoder is in state *a* and the input is **0**, the output is **00** (solid line), and the encoder remains in state *a*. Note that the encoder cannot go directly from state *a* to states *c* or *d*. From any given state, the encoder can go to only two states directly by inputting a single data bit. This is an extremely important observation, which will be used later. The encoder goes from state *a* to state *b* (when the input is **1**), or to state *a* (when the input is **0**), and so on. The encoder cannot go from *a* to *c* in one step. It must go from *a* to *b* to *c*, or from *a* to *b* to *d* to *c*, and so on. We can also verify these facts from the code tree. Figure 16.7b contains the complete information of the code tree.

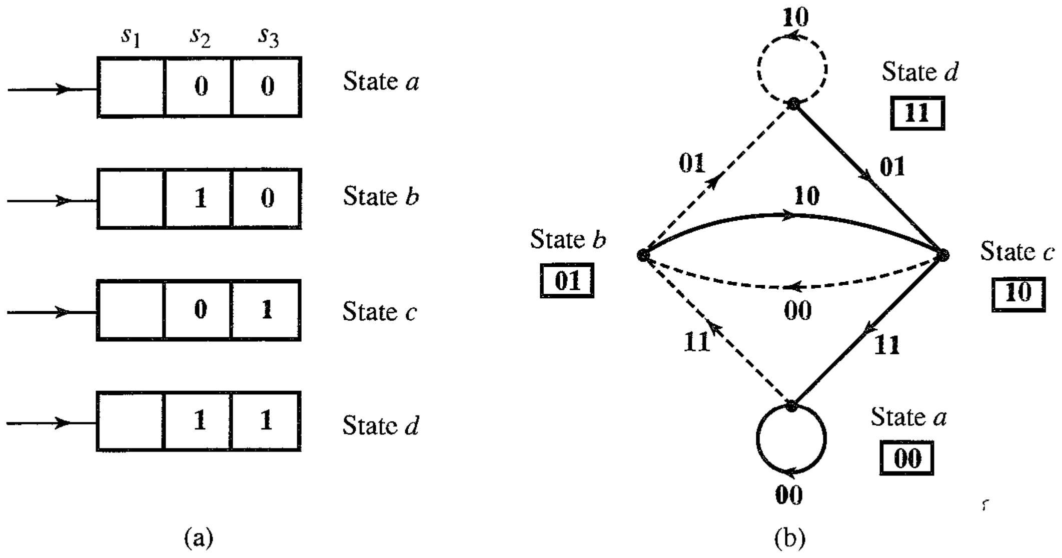


Figure 16.7 State diagram of the coder in Fig. 16.5.

Another useful way of representing the code tree is the trellis diagram in Fig. 16.8. The diagram starts from scratch (all **0**'s in the shift register, that is, state *a*) and makes transitions corresponding to each input data digit. These transitions are denoted by a solid line for the next data digit **0** and by a dashed line for the next data digit **1**. Thus, when the first input digit is **0**, the encoder output is **00** (solid line), and when the input digit is **1**, the encoder output is **11** (dashed line). This is readily seen from Fig. 16.7b. We continue this way with the second input digit. After the first two input digits, the encoder is in one of the four states *a*, *b*, *c*, or *d*, as shown in Fig. 16.8. If the encoder is in state *a* (previous two data digits **00**), it goes to state *b* if the next input bit is **1** or remains in state *a* if the next input bit is **0**. In so doing, the encoder output is **11** (*a* to *b*) or **00** (*a* to *a*). Note that the structure of the trellis diagram is completely repetitive, as expected, and can be readily drawn using the state diagram in Fig. 16.7b.

Decoding

We shall consider two important techniques: (1) maximum-likelihood decoding (Viterbi's algorithm) and (2) sequential decoding.

Maximum-Likelihood Decoding—Viterbi's Algorithm: Among various decoding methods for convolutional codes, Viterbi's maximum-likelihood algorithm[4] is one of the best techniques yet evolved for digital communications where energy efficiency dominates in importance. It permits major equipment simplification while obtaining the full performance benefits of maximum-likelihood decoding. The decoder structure is relatively simple for short constraint lengths, making decoding feasible at relatively high rates of up to 100 Mbit/s.

The maximum-likelihood receiver implies selecting a code word closest to the received word. Because there are 2^k code words, the maximum-likelihood decision involves storage of 2^k words and their comparison with the received word. This calculation is extremely difficult for large k and would result in an overly complex decoder.

A major simplification was made by Viterbi in the maximum-likelihood calculation by noting that each of the four nodes (*a*, *b*, *c*, and *d*) has only two predecessors; that is, each node

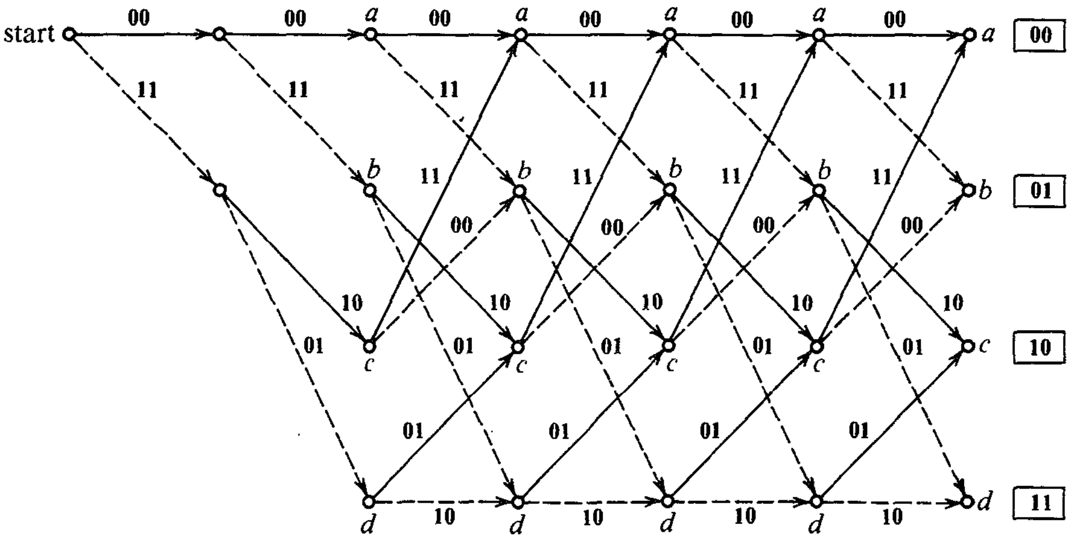


Figure 16.8 Trellis diagram for the coder in Fig. 16.5.

can be reached through two nodes only (see Fig. 16.7 or 16.8), and only the path that agrees most with the received sequence (the minimum-distance path) need be retained for each node. To understand this, consider the trellis diagram in Fig. 16.8. Our problem is as follows: Given a received sequence of bits, we need to find a path in the trellis diagram with the output digit sequence that agrees most with the received sequence.

Suppose the first six received digits are **01 00 01**. We shall consider two paths of three branches (for six digits) leading to each of the nodes a, b, c, and d. Out of the two paths reaching into each node, we shall retain only the one that agrees most with the received sequence **01 00 01** (the minimum-distance path). The retained path is called the **survivor** at that node. There are two paths (**00 00 00** and **11 10 11**) that arrive at the third-level node a. These paths are at distances of 2 and 3, respectively, from the received sequence **01 00 01**. Hence, the survivor at the third-level node a is **00 00 00**. We repeat the procedure for nodes b, c, and d. For example, the two paths reaching the third-level node c (the node after three branches) are **00 11 10** and **11 01 01**, at distances of 5 and 2, respectively, from the received sequence **01 00 01**. Hence, the survivor at the third-level node c is **11 01 01**. Similarly, we find survivors at the third-level nodes b and d. With four paths eliminated, the four survivor paths are the only contenders. The reason behind eliminating the other four paths is as follows. The two paths merging at the third-level node a, for example, imply that the previous two data digits are identical (viz., **00**). Hence, regardless of what the future data digits are, both paths must merge at this node a and follow a common path in the future. Clearly, the survivor path is the minimum-distance path between the two, regardless of future data digits. What we need to remember is the four survivor paths and their distances from the received sequence. In general, the number of survivor paths is equal to the number of states, that is, 2^{N-1}.

Once we have survivors at all the third-level nodes, we look at the next two received digits. Suppose these are **00** (i.e., the received sequence is **01 00 01 00**). We now compare the two survivors that merge into the fourth-level node a. These are the survivors at nodes a and c of the third level, with paths **00 00 00 00** and **11 01 01 11**, respectively, and distances of 2 and 4 from the received sequence **01 00 01 00**. Hence, the path **00 00 00 00** is the survivor at the fourth-level node a. We repeat this procedure for nodes b, c, and d and continue in this manner until the end. Note that only two paths merge in a node and there are only four contending paths (the four survivors at nodes a, b, c, and d) until the end. The only remaining problem is how to truncate the algorithm and ultimately decide on one path rather than four. This is done by forcing the last two data digits to be **00** (dummy or augmented data). When the first dummy **0** enters the register, we consider the survivors only at nodes a and c. The survivors at nodes b and d are discarded because these nodes can be reached only when the input bit is **1**, as seen from the state or trellis diagram. When the second dummy **0** enters the register, we consider only the survivor at node a. We discard the survivor at node c because the last two dummy data bits **00** lead the encoder to state a. In terms of the trellis diagram, this means that the number of states is reduced from four to two (a and c) by insertion of the first zero and to a single state (a) by insertion of the second zero.

With the Viterbi algorithm, storage and computational complexity are proportional to 2^N and are very attractive for constraint length $N < 10$. To achieve very low error probabilities, longer constraint lengths are required, and sequential decoding (to be discussed next) becomes attractive.

Sequential Decoding: In this technique, proposed by Wozencraft, the decoder complexity increases linearly rather than exponentially. To explain this technique, let us consider a coder with $N = 4$ and $v = 3$ (Fig. 16.9). The code tree for this coder is shown in Fig. 16.10. Each data digit generates three ($v = 3$) output digits but affects four groups of three digits (12 digits) in all.

In this decoding scheme, we observe only three (or v) digits at a time to make a tentative decision, with readiness to change our decision if it creates difficulties later. A sequential detector acts much like a driver who occasionally makes a wrong choice at a fork in the road, but quickly discovers the error (because of road signs), goes back, and tries the other path.

Applying this insight to our decoding problem, the analogous procedure would be as follows. We look at the first three received digits. There are only two paths of three digits from

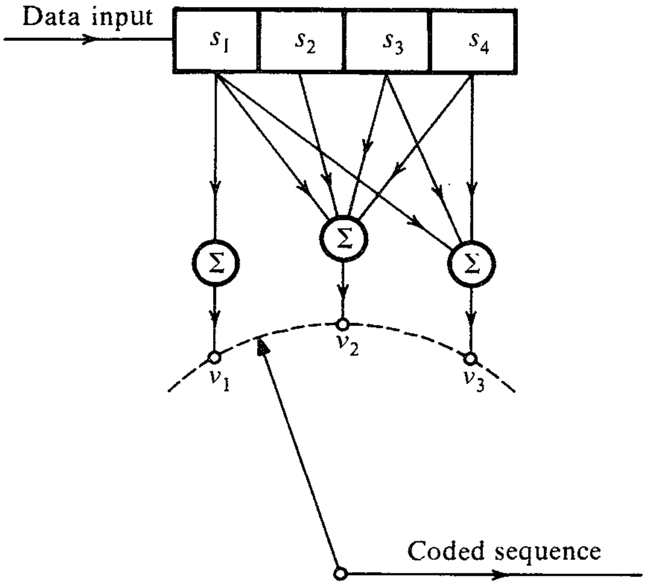


Figure 16.9 Convolution coder.

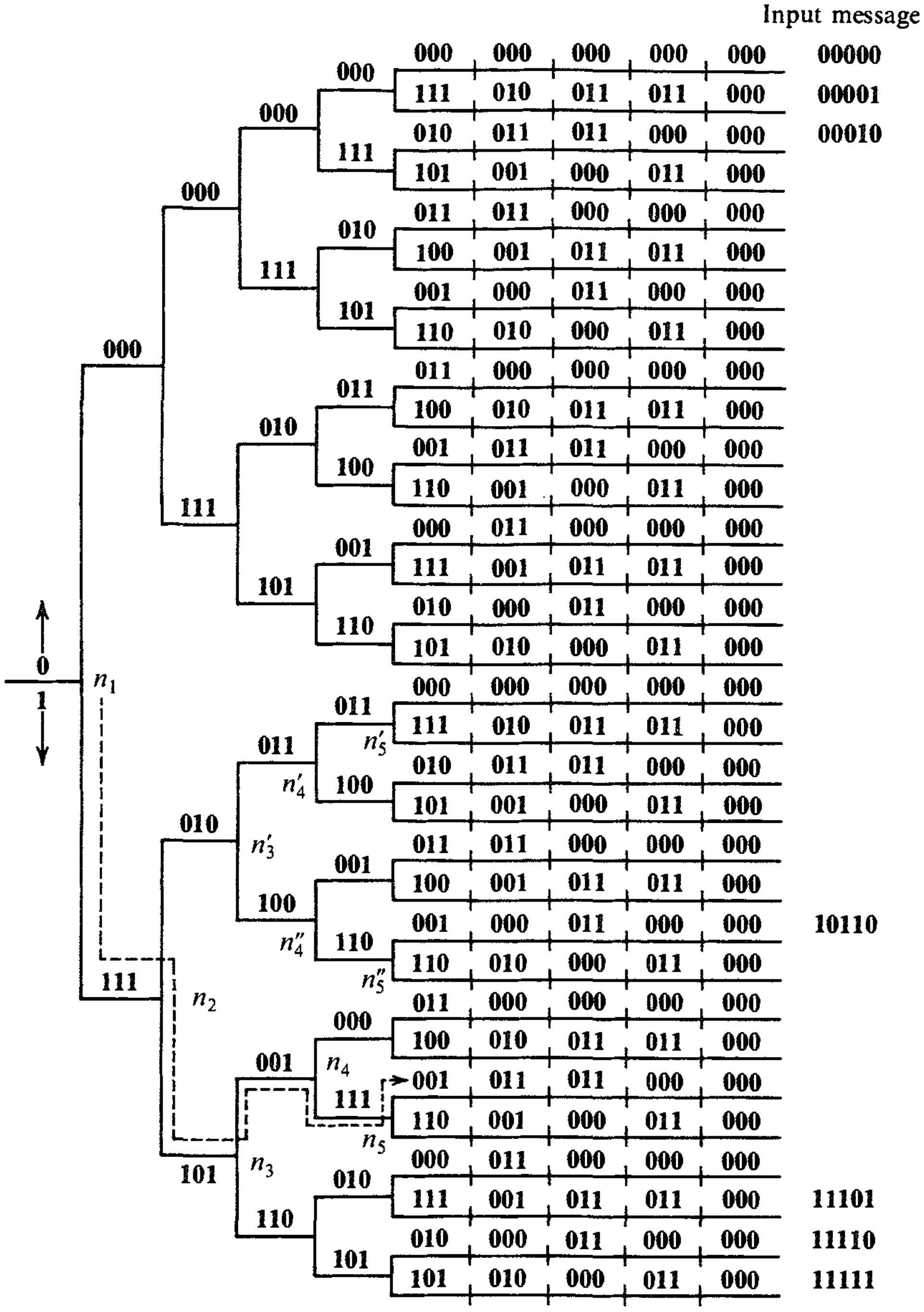

Input message
000 000 000 000 000 00000
111 010 011 011 000 00001
010 011 011 000 000 00010
101 001 000 011 000
011 011 000 000 000
100 001 011 011 000
001 000 011 000 000
110 010 000 011 000
011 000 000 000 000
100 010 011 011 000
001 011 011 000 000
110 001 000 011 000
000 011 000 000 000
111 001 011 011 000
010 000 011 000 000
101 010 000 011 000
000 000 000 000 000
111 010 011 011 000
010 011 011 000 000
101 001 000 011 000
011 011 000 000 000
100 001 011 011 000
001 000 011 000 000 10110
110 010 000 011 000
011 000 000 000 000
100 010 011 011 000
001 011 011 000 000
110 001 000 011 000
000 011 000 000 000
111 001 011 011 000 11101
010 000 011 000 000 11110
101 010 000 011 000 11111
000
000
111
000
010
111
101
000
011
010
100
111
001
101
110
011
011
100
010
001
100
110
111
000
001
111
101
010
110
101
0
1
n_1
n_2
n_3
n_4
n_5
n'_3
n'_4
n'_5
n''_4
n''_5

Figure 16.10 Code tree for the coder in Fig. 16.9.

the initial node n_1. We choose that path whose sequence is at the shortest Hamming distance from the first three received digits. We thus progress to the most likely node. From this node there are two paths of three digits. We look at the second group of the three received digits and choose that path whose sequence is closest to these received digits. We progress this way until the fourth node. If we were unlucky enough to have a large number of errors in a certain received group of v digits, we will take a wrong turn, and from there on we will find it more difficult to match the received digits with those along the paths available from the wrong node. This is the clue to the realization that an error has been made. Let us explain this by an example.

Suppose a data sequence **11010** is encoded by the coder in Fig. 16.9. Because $N = 4$, we add three dummy **0**'s to this sequence so that the augmented data sequence is **11010000**. The coded sequence will be (see the code tree in Fig. 16.10) **111 101 001 111 001 011 011 000**. Let the received sequence be **101 011 001 111 001 011 011 000** (three errors: one in the first group and two in the second group). We start at the initial node n_1. The first received group **101** (one error) being closer to **111**, we make a correct decision to go to node n_2. But the second group **001** (two errors) is closer to **010** than to **101** and will lead us to the wrong node n_3' rather than to n_3. From here on we are on the wrong track, and, hence, the received digits will not match any path starting from n_3'. The third received group is **001** and does not match any sequence starting at n_3' (viz., **001** and **100**). But it is closer to **011**. Hence, we go to node n_4'. Here again the fourth received group **111** does not match any group starting at n_4' (viz., **011** and **100**). But it is closer to **011**. This takes us to node n_5'. It can be seen that the Hamming distance between the sequence of 12 digits along the path $n_1n_2n_3'n_4'n_5'$ and the first 12 received digits is 4, indicating four errors in 12 digits (if our path is correct). Such a high number of errors should immediately make us suspicious. If P_e is the digit error probability, then the expected number of errors n_e in d digits is P_ed. Because P_e is on the order of 10^{-4} to 10^{-6}, four errors in 12 digits is unreasonable. Hence, we go back to node n_3' and try the lower branch, leading to n_5''. This path $n_1n_2n_3'n_4''n_5''$ is even worse than the previous one, because it gives five errors in 12 digits. Hence, we go back even farther to node n_2 and try the path leading to n_3 and farther. We find the path $n_1n_2n_3n_4n_5$, giving three errors. If we go back still farther to n_1 and try alternate paths, we find that none yields less than five errors. Thus, the correct path is taken as $n_1n_2n_3n_4n_5$, giving three errors. If we go back still farther to n_1 and try alternate paths, we find that none yields less than five errors. Thus, the correct path is taken as $n_1n_2n_3n_4n_5$. This enables us to decode the first transmitted digit as **1**. Next, we start at node n_2, discard the first three received digits, and repeat the procedure to decode the second transmitted digit. We repeat this until all the digits are decoded.

The next important question concerns the criterion for deciding when the wrong path is chosen. The plot of the expected number of errors n_e as a function of the number of decoded digits d is a straight line ($n_e = P_ed$) with slope P_e, as shown in Fig. 16.11. The actual number of errors along the path is also plotted. If the errors remain within a limit (the discard level), the decoding continues. If at some point it is found that the errors exceed the discard level, we go back to the nearest decision node and try an alternate path. If errors still increase beyond the discard level, we then go back one more node along the path and try an alternate path. The process continues until the errors are within the set limit. By making the discard level very stringent (close to the expected error curve), we reduce the average number of computations. On the other hand, if the discard level is made too stringent, the decoder will discard all possible paths in some extremely rare cases where the noise may cause an unusually large number of errors. This difficulty is usually resolved by starting with a stringent discard level. If on rare occasions the decoder rejects all paths, the discard level can be relaxed bit by bit until one of the paths is acceptable.

It can be shown that the error probability in this scheme decreases exponentially as N, whereas the system complexity grows only linearly with k. The efficiency is $\eta \simeq 1/v$. It can be

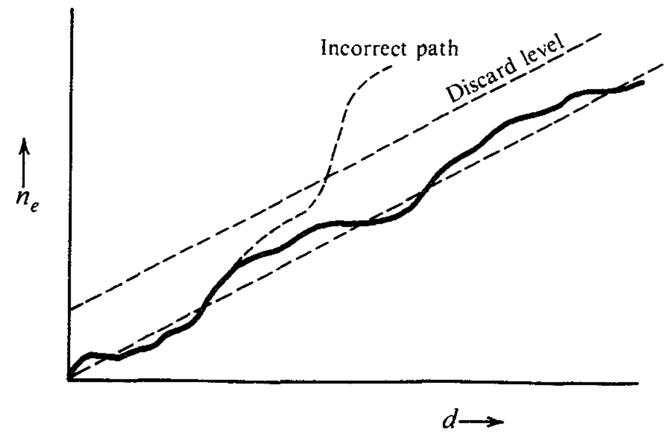


Figure 16.11 Setting the threshold in sequential decoding.

shown that for $\eta < \eta_o$, the average number of incorrect branches searched per decoded digit is bounded, whereas for $\eta > \eta_o$ it is not; hence η_o is called the **computational cutoff rate.**

There are several disadvantages to sequential decoding:

1. The number of incorrect path branches, and consequently the computation complexity, is a random variable depending on the channel noise.
2. To make storage requirements easier, the decoding speed has to be maintained at 10 to 20 times faster than the incoming data rate. This limits the maximum data rate capability.
3. The average number of branches can occasionally become very large and may result in a storage overflow, causing relatively long sequences to be erased.

A third technique for decoding convolutional codes is **feedback decoding**, with threshold decoding[5] as a subclass. Threshold decoders are easily implemented. Their performance, however, does not compare favorably with the previous two methods.

16.7 COMPARISON OF CODED AND UNCODED SYSTEMS

It is instructive to compare the error probability of coded and uncoded schemes under similar constraints of power and information rate.

Let us consider a t-error correcting (n, k) code. In this case, k information digits are coded into n digits. For a proper comparison, we shall assume that k information digits are transmitted in the same time interval over both systems and that the transmitted power S_i is also maintained the same for both systems. Because only k digits are required to be transmitted over the uncoded system (versus n over the coded one), the bit rate R_b is lower for the uncoded system by a factor of k/n as compared to the coded system. Hence, for a given power, E_b is higher for the uncoded case as compared to the coded case by a factor n/k. This tends to reduce the bit error probability for the uncoded case. Let P_{eu} and P_{ec} represent the digit error probabilities in the uncoded and coded cases, respectively.

For the uncoded case, a word of k digits will be received wrong if any one of the k digits is in error. If P_{Eu} and P_{Ec} represent the word error probabilities of the uncoded and coded systems, respectively, then

$$
\begin{aligned}
P_{Eu} &= 1 - P(\text{all } k \text{ digits received correct}) \\
&= 1 - (1 - P_{eu})^k && (16.21a) \\
&\simeq kP_{eu} \qquad P_{eu} \ll 1 && (16.21b)
\end{aligned}
$$

For a t-error correcting (n, k) code, the received word will be in error if more than t errors occur in n digits. If $P(j, n)$ is the probability of j errors in n digits, then

$$P_{Ec} = \sum_{j=t+1}^{n} P(j, n)$$

Because there are $\binom{n}{j}$ ways in which j errors can occur in n digits (Example 10.6),

$$P(j, n) = \binom{n}{j}(P_{ec})^j(1 - P_{ec})^{n-j}$$

and

$$P_{Ec} = \sum_{j=t+1}^{n} \binom{n}{j}(P_{ec})^j(1 - P_{ec})^{n-j} \qquad (16.22)$$

For $P_{Ec} \ll 1$, the first term in the summation in Eq. (16.22) dominates all the other terms, and we are justified in ignoring all but the first term. Hence,

$$
\begin{aligned}
P_{Ec} &= \binom{n}{t+1}(P_{ec})^{t+1}(1 - P_{ec})^{n-(t+1)} && (16.23a) \\
&\simeq \binom{n}{t+1}(P_{ec})^{t+1} \qquad P_{ec} \ll 1 && (16.23b)
\end{aligned}
$$

For further comparison, we must assume some specific transmission scheme. Let us consider a coherent PSK scheme. In this case for an AWGN channel,

$$P_{eu} = Q\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right) \qquad (16.24a)$$

and because E_b for the coded case is k/n times that for the uncoded case,

$$P_{ec} = Q\left(\sqrt{\frac{2kE_b}{n\mathcal{N}}}\right) \qquad (16.24b)$$

Hence,

$$P_{Eu} = kQ\left(\sqrt{\frac{2E_b}{\mathcal{N}}}\right) \qquad (16.25a)$$

$$P_{Ec} = \binom{n}{t+1}\left[Q\left(\sqrt{\frac{2kE_b}{n\mathcal{N}}}\right)\right]^{t+1} \qquad P_{ec} \ll 1 \qquad (16.25b)$$

To compare coded and uncoded systems, we could plot P_{Eu} and P_{Ec} as functions of $E_b/\mathcal{N}$. Because Eqs. (16.25) involve parameters t, n, and k, a proper comparison requires families

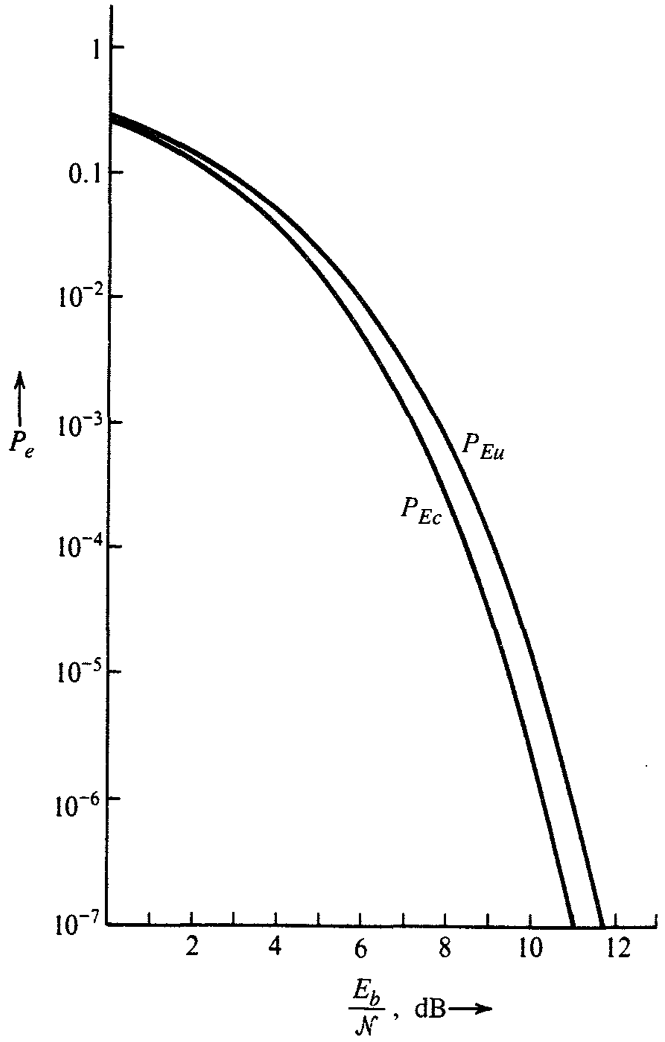


Figure 16.12 Comparison of coded and uncoded systems.

of plots. For the case of a (7, 4) single-error correcting code ($t = 1, n = 7$, and $k = 4$), P_{Ec} and P_{Eu} in Eqs. (16.25) are plotted in Fig. 16.12 as a function of $E_b/\mathcal{N}$. Observe that the coded scheme is superior to the uncoded scheme, but the improvement (about 1 dB) is not too significant. For large n and k, however, the coded scheme can become significantly superior to the uncoded one. For practical channels plagued by fading and impulse noise, coding can yield substantial gains.

EXAMPLE 16.6 Compare the performance of an AWGN BSC using a single-error correcting (15, 11) code with that of the same system using uncoded transmission, given that $E_b/\mathcal{N} = 9.12$ for the uncoded scheme and coherent PSK is used to transmit the data.

From Eq. (16.25a),

$$P_{Eu} = 11Q(\sqrt{18.24}) = 1.1 \times 10^{-4}$$

and from Eq. (16.25b),

$$P_{Ec} = \binom{15}{2}\left[Q\left(\sqrt{\frac{11}{15}(18.24)}\right)\right]^2$$

$$= 105(1.37 \times 10^{-4})^2 = 1.96 \times 10^{-6}$$

Note that the word error probability of the coded system is reduced by a factor of 56. On the other hand, if we wish to achieve the error probability of the coded transmission (1.96×10^{-6}) using the uncoded system, we must increase the transmitted power. If E_b' is the new value of E_b to achieve $P_{Eu} = 1.96 \times 10^{-6}$,

$$P_{Eu} = 11Q\left(\sqrt{\frac{2E_b'}{\mathcal{N}}}\right) = 1.96 \times 10^{-6}$$

This gives $E_b'/\mathcal{N} = 10.7$. This is an increase over the old value of 9.12 by a factor of 1.17, or 0.7 dB.

REFERENCES

1. W. W. Peterson and E. J. Weldon, Jr., *Error Correcting Codes*, 2nd ed., Wiley, New York, 1972.
2. S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Englewood-Cliffs, NJ, 1983.
3. P. Elias, "Coding for Noisy Channels," *IRE Nat. Conv. Rec.*, vol. 3, part 4, pp. 37–46, 1955.
4. A. J. Viterbi, "Convolutional Codes and Their Performance in Communications Systems," *IEEE Trans. Commun. Technol.*, vol. CT-19, pp. 751–771, Oct. 1971.
5. J. L. Massey, *Threshold Decoding*, M.I.T. Press, Cambridge, MA, 1963.

PROBLEMS

16.1-1 Golay's (23, 12) codes are three-error correcting codes. Verify that $n = 23$ and $k = 12$ satisfies the Hamming bound exactly for $t = 3$.

16.1-2 (a) Determine the Hamming bound for a ternary code.

(b) A ternary (11, 6) code exists that can correct up to two errors. Verify that this code satisfies the Hamming bound exactly.

16.1-3 Confirm the possibility of a (18, 7) binary code that can correct up to three errors. Can this code correct up to four errors?

16.2-1 If $\boldsymbol{G}$ and $\boldsymbol{H}$ are the generator and parity-check matrices, respectively, then show that

$$\boldsymbol{G}\boldsymbol{H}^T = 0$$

16.2-2 Given a generator matrix

$$\boldsymbol{G} = [1 \quad 1 \quad 1]$$

construct a (3, 1) code. How many errors can this code correct? Find the code word for data vectors $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{d} = \boldsymbol{1}$. Comment.

16.2-3 Repeat Prob. 16.2-2 for

$$G = [1 \quad 1 \quad 1 \quad 1 \quad 1]$$

This gives a (5, 1) code.

16.2-4 Suppose we wish to increase reliability by repeating a message three times, e.g., by transmitting data **0** by **000** and **1** by **111**. This is a (3, 1) code.

(a) Is this a systematic code?

(b) If so, find the generating matrix G.

16.2-5 Consider the following $(k+1, k)$ systematic linear block code with the parity-check digit c_{k+1} given by

$$c_{k+1} = d_1 + d_2 + \cdots + d_k$$

(a) Construct the appropriate generator matrix for this code.

(b) Construct the code generated by this matrix for $k = 3$.

(c) Determine the error detecting or correcting capabilities of this code.

(d) Show that

$$cH^T = 0$$

and

$$rH^T = \begin{cases} \mathbf{0} & \text{if no error occurs} \\ \mathbf{1} & \text{if single error occurs} \end{cases}$$

16.2-6 Consider a generator matrix G for a nonsystematic (6, 3) code:

$$G = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Construct the code for this G, and show that $d_{\min}$, the minimum distance between code words, is 3. Consequently, this code can correct at least one error.

16.2-7 Repeat Prob. 16.2-6 if

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

16.2-8 Find a generator matrix G for a (15, 11) single-error correcting linear block code. Find the code word for the data vector **10111010101**.

16.2-9 For a (6, 3) systematic linear block code, the three parity-check digits c_4, c_5, and c_6 are

$$c_4 = d_1 + d_2 + d_3$$
$$c_5 = d_1 + d_2$$
$$c_6 = d_1 + d_3$$

(a) Construct the appropriate generator matrix for this code.

(b) Construct the code generated by this matrix.

(c) Determine the error correcting capabilities of this code.

(d) Prepare a suitable decoding table.

(e) Decode the following received words: **101100**, **000110**, **101010**.

16.2-10 (a) Construct a code table for the (6, 3) code generated by the matrix $\boldsymbol{G}$ in Prob. 16.2-6.

(b) Prepare a suitable decoding table.

16.2-11 Construct a single-error correcting (7, 4) linear block code (Hamming code) and the corresponding decoding table.

16.2-12 For the (6, 3) code in Example 16.1, the decoding table is Table 16.3. Show that if we use this decoding table, and a two-error pattern **010100** or **001001** occurs, it will not be corrected. If it is desired to correct a single two-error pattern **010100** (along with six single-error patterns), construct the appropriate decoding table and verify that it does indeed correct one two-error pattern **010100** and that it cannot correct any other two-error patterns.

16.2-13 (a) Given $k = 8$, find the minimum value of n for a code that can correct at least one error.

(b) Choose a generator matrix $\boldsymbol{G}$ for this code.

(c) How many double errors can this code correct?

(d) Construct a decoding table (syndromes and corresponding correctable error patterns).

16.2-14 Consider a (6, 2) code generated by the matrix

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}$$

(a) Construct the code table for this code and determine the minimum distance between code words.

(b) Prepare a suitable decoding table. *Hint:* This code can correct all single-error patterns, seven double-error patterns, and two triple-error patterns. Choose the desired seven double-error patterns and the two triple-error patterns.

16.3-1 (a) Construct a systematic (7, 4) cyclic code using the generator polynomial $g(x) = x^3 + x + 1$.

(b) What are the error correcting capabilities of this code?

(c) Construct the decoding table.

(d) If the received word is **1101100**, determine the transmitted data word.

16.3-2 A three-error correcting (23, 12) Golay code is a cyclic code with a generator polynomial

$$g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$$

Determine the code words for the data vectors **000011110000**, **101010101010**, and **110001010111 10**.

16.3-3 Factorize the polynomial

$$x^3 + x^2 + x + 1$$

Hint: A third-order polynomial must have one factor of first order. The only first-order polynomials that are prime (not factorable) are x and $x+1$. Since x is not a factor of the given polynomial, try $x+1$. Divide x^3+x^2+x+1 by $x+1$.

16.3-4 The concept explained in Prob. 16.3-3 can be extended to factorize any higher order polynomial. Using this technique, factorize

$$x^5+x^4+x^2+1$$

Hint: There must be at least one first-order factor. Try dividing by the two first-order prime polynomials x and $x+1$. The given fifth-order polynomial can now be expressed as $\phi_1(x)\phi_4(x)$, where $\phi_1(x)$ is a first-order polynomial and $\phi_4(x)$ is a fourth-order polynomial which may or may not contain a first-order factor. Try dividing $\phi_4(x)$ by x and $x+1$. If it does not work, it must have two second-order polynomials both of which are prime. The possible second-order polynomials are x^2, x^2+1, x^2+x, and x^2+x+1. Determine which of these are prime (not divisible by x or $x+1$). Now try dividing $\phi_4(x)$ by these prime polynomials of the second order. If neither divides, $\phi_4(x)$ must be a prime polynomial of the fourth order and the factors are $\phi_1(x)$ and $\phi_4(x)$.

16.3-5 Use the concept explained in Prob. 16.3-4 to factorize a seventh-order polynomial x^7+1. *Hint*: Determine prime factors of first-, second-, and third-order polynomials. The possible third-order polynomials are x^3, x^3+1, x^3+x, x^3+x+1, x^3+x^2, x^3+x^2+1, x^3+x^2+x, and x^3+x^2+x+1. See hint in Prob. 16.3-4.

16.3-6 Equation (16.15) suggests a method of constructing a generator matrix $\boldsymbol{G}'$ for a cyclic code,

$$\boldsymbol{G}' = \begin{bmatrix} x^{k-1}g(x) \\ x^{k-2}g(x) \\ \cdots \\ g(x) \end{bmatrix} = \begin{bmatrix} g_1 & g_2 & \cdots & g_{n-k+1} & 0 & 0 & \cdots & 0 \\ 0 & g_1 & g_2 & \cdots & g_{n-k+1} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & g_1 & g_2 & \cdots & g_{n-k+1} \end{bmatrix}$$

where $g(x) = g_1x^{n-k} + g_2x^{n-k-1} + \cdots + g_{n-k+1}$ is the generator polynomial. This is, in general, a nonsystematic cyclic code.

(a) For a single-error correcting (7, 4) cyclic code with a generator polynomial $g(x) = x^3+x^2+1$, find $\boldsymbol{G}'$ and construct the code.

(b) Verify that this code is identical to that derived in Example 16.3 (Table 16.4).

16.3-7 The generator matrix $\boldsymbol{G}$ for a systematic cyclic code (see Prob. 16.3-6) can be obtained by realizing that adding any row of a generator matrix to any other row yields another valid generator matrix, because the code word is formed by linear combinations of data digits. Also, a generator matrix for a systematic code must have an identity matrix $\boldsymbol{I}_k$ in the first k columns. Such a matrix is formed step by step as follows. Observe that each row in $\boldsymbol{G}'$ in Prob. 16.3-6 is a left shift of the row below it, with the last row being $g(x)$. Start with the kth (last) row $g(x)$. Because $g(x)$ is of the order $n-k$, this row has the element 1 in the kth column, as required. For the $(k-1)$th row, use the last row with one left shift. We require a 0 in the kth column of the $(k-1)$th row to form $\boldsymbol{I}_k$. If there is a 0 in the kth column of this $(k-1)$th row, we accept it as a valid $(k-1)$th row. If not, then we add the kth row to the $(k-1)$th row to obtain 0 in its kth column. The resulting row is the final $(k-1)$th row. This row with a single left shift serves as the $(k-2)$th row. But if this newly formed $(k-2)$th row does not have a 0 in its kth column, we add the kth (last) row to it to get the desired 0. We continue this way until all k rows are formed. This gives the generator matrix for a systematic (n, k) cyclic code.

(a) For a single-error correcting (7, 4) systematic cyclic code with a generator polynomial $g(x) = x^3 + x^2 + 1$, find $\boldsymbol{G}$ and construct the code.

(b) Verify that this code is identical to that in Table 16.5 (Example 16.4).

16.3-8 (a) Find the generator matrix $\boldsymbol{G}'$ for a nonsystematic (7, 4) cyclic code using the generator polynomial $g(x) = x^3 + x + 1$.

(b) Find the code generated by this matrix $\boldsymbol{G}'$.

(c) Determine the error correcting capabilities of this code.

16.3-9 Find the generator matrix $\boldsymbol{G}$ for a systematic (7, 4) cyclic code using the generator polynomial $g(x) = x^3 + x + 1$ (see Prob. 16.3-7).

16.4-1 The simple burst-error detecting code in Fig. 16.3 can also be used as a single-error correcting code with a slight modification. The k data digits are divided into groups of b digits in length, as in Fig. 16.3. To each group we add one parity-check digit, so that each segment now has $b + 1$ digits (b data digits and one parity-check digit). The parity-check digit is chosen to ensure that the total number of **1**'s in each segment of $b + 1$ digits is even. Now we consider these digits as our new data and augment them with the last segment of $b + 1$ parity-check digits, as was done in Fig. 16.3. The data in Fig. 16.3 will be transmitted thus:

10111 01010 11011 10001 11000 01111

Show that this (30, 20) code is capable of single.error correction as well as the detection of a single burst of length 5.

16.5-1 Discuss the error correcting capabilities of an interlaced (λn, λk) cyclic code with $\lambda = 10$ and using a three-error correcting (31, 16) BCH code.

16.6-1 For the convolutional encoder in Fig. 16.5, the received bits are **01 00 01 00 10 11 11 00**. Decode this sequence using Viterbi's algorithm and the trellis diagram in Fig. 16.8.

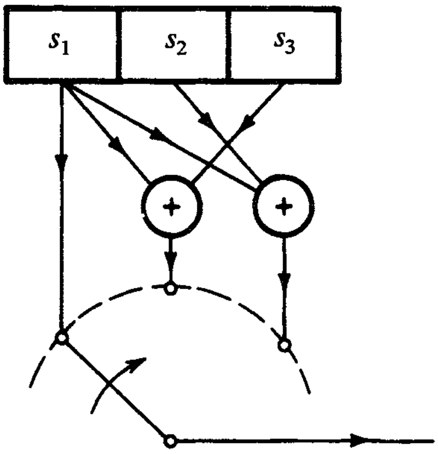


Figure P16.6-2

16.6-2 For the convolutional encoder shown in Fig. P16.6-2:

(a) Draw the state and trellis diagrams and determine the output digit sequence for the data digits **11010100**.

(b) Use Viterbi's algorithm to decode the following received sequences:

(i) 100 110 111 101 001 101 001 010

(ii) 010 110 111 101 101 101 001 010

(iii) 111 110 111 111 001 101 001 101

16.7-1 Uncoded data is transmitted using PSK over an AWGN channel with $E_b/\mathcal{N} = 9.12$. This data is now coded using a three-error correcting (23, 12) Golay code (see Prob. 16.1-1) and transmitted over the same channel at the same data rate and with the same transmitted power.

(a) Determine the word error probabilities P_{Eu} and P_{Ec} for the coded and the uncoded systems.

(b) If it is decided to achieve the error probability P_{Ec} computed in part **(a)** using the uncoded system by increasing the transmitted power, determine the required value of $E_b/\mathcal{N}$.

Appendix A

ORTHOGONALITY OF SOME SIGNAL SETS

A.1 Orthogonality of the Trigonometric and Exponential Signal Set

Consider an integral I defined by

$$I = \int_{T_0} \cos n\omega_0 t \, \cos m\omega_0 t \, dt \qquad (A.1a)$$

where $\int_{T_0}$ stands for integration over any contiguous interval of $T_0 = 2\pi/\omega_0$ seconds. By using a trigonometric identity (see Appendix D), Eq. (A.1a) can be expressed as

$$I = \frac{1}{2}\left[\int_{T_0} \cos\,(n+m)\omega_0 t \, dt + \int_{T_0} \cos\,(n-m)\omega_0 t \, dt\right] \qquad \text{(A.1b)}$$

Since cos $\omega_0 t$ executes one complete cycle during any interval of T_0 seconds, cos $(n+m)\omega_0 t$ executes $(n+m)$ complete cycles during any interval of duration T_0. Therefore, the first integral in Eq. (A.1b), which represents the area under $(n+m)$ complete cycles of a sinusoid, equals zero. The same argument shows that the second integral in Eq. (A.1b) is also zero, except when $n = m$. Hence, I in Eq. (A.1b) is zero for all $n \neq m$. When $n = m$, the first integral in Eq. (A.1b) is still zero, but the second integral yields

$$I = \frac{1}{2}\int_{T_0} dt = \frac{T_0}{2}$$

Thus,

$$\int_{T_0} \cos n\omega_0 t \, \cos m\omega_0 t \, dt = \begin{cases} 0 & n \neq m \\ \dfrac{T_0}{2} & m = n \neq 0 \end{cases} \qquad \text{(A.2a)}$$

Using similar arguments, we can show that

$$\int_{T_0} \sin n\omega_0 t \, \sin m\omega_0 t \, dt = \begin{cases} 0 & n \neq m \\ \dfrac{T_0}{2} & n = m \neq 0 \end{cases} \qquad \text{(A.2b)}$$

and

$$\int_{T_0} \sin n\omega_0 t \cos m\omega_0 t \, dt = 0 \qquad \text{all } n \text{ and } m \tag{A.2c}$$

A.2 Orthogonality of the Exponential Signal Set

The set of exponentials $e^{jn\omega_0 t}$ ($n = 0, \pm 1, \pm 2, \dots$) is orthogonal over any interval of duration T_0, that is,

$$\int_{T_0} e^{jm\omega_0 t}(e^{jn\omega_0 t})^* \, dt = \int_{T_0} e^{j(m-n)\omega_0 t} \, dt = \begin{cases} 0 & m \neq n \\ T_0 & m = n \end{cases} \tag{A.3}$$

Let the integral on the left-hand side of Eq. (A.3) be I,

$$\begin{aligned} I &= \int_{T_0} e^{jm\omega_0 t}(e^{jn\omega_0 t})^* \, dt \\ &= \int_{T_0} e^{j(m-n)\omega_0 t} \, dt \end{aligned} \tag{A.4}$$

The case $m = n$ is trivial. In this case the integrand is unity, and $I = T_0$. When $m \neq n$,

$$\begin{aligned} I &= \frac{1}{j(m-n)\omega_0} e^{j(m-n)\omega_0 t} \Big|_{t_1}^{t_1+T_0} \\ &= \frac{1}{j(m-n)\omega_0} e^{j(m-n)\omega_0 t_1} [e^{j(m-n)\omega_0 T_0} - 1] = 0 \end{aligned}$$

The last result follows from the fact that $\omega_0 T_0 = 2\pi$, and $e^{j2\pi k} = 1$ for all integral values of k.

Appendix B

SCHWARZ INEQUALITY

Prove the following Schwarz inequality for a pair of real finite energy signals $f(t)$ and $g(t)$:

$$\left[\int_a^b f(t)g(t)\,dt\right]^2 \leq \left[\int_a^b f^2(t)\,dt\right]\left[\int_a^b g^2(t)\,dt\right] \tag{B.1}$$

with equality only if $g(t) = cf(t)$, where c is an arbitrary constant.

The Schwarz inequality for finite-energy, complex-valued functions $X(\omega)$ and $Y(\omega)$ is given by

$$\left|\int_{-\infty}^{\infty} X(\omega)Y(\omega)\,d\omega\right|^2 \leq \int_{-\infty}^{\infty} |X(\omega)|^2 d\omega \int_{-\infty}^{\infty} |Y(\omega)|^2 d\omega \tag{B.2}$$

with equality only if $Y(\omega) = cX^*(\omega)$, where c is an arbitrary constant.

We can prove Eq. (B.1) as follows: For any real value of λ, we know that

$$\int_a^b [\lambda f(t) - g(t)]^2 dt \geq 0$$

or

$$\lambda^2 \int_a^b f^2(t)\,dt - 2\lambda \int_a^b f(t)g(t)\,dt + \int_a^b g^2(t)\,dt \geq 0$$

Because this quadratic in λ is nonnegative for any value of λ, its discriminant must be nonpositive, and Eq. (B.1) follows. If the discriminant equals zero, then for some value of $\lambda = c$, the quadratic equals zero. This is possible only if $cf(t) - g(t) = 0$, and the result follows.

To prove Eq. (B.2), we observe that $|X(\omega)|$ and $|Y(\omega)|$ are real functions and inequality B.1 applies. Hence,

$$\left[\int_a^b |X(\omega)Y(\omega)|\,d\omega\right]^2 \leq \int_a^b |X(\omega)|^2 d\omega \int_a^b |Y(\omega)|^2 d\omega \tag{B.3}$$

with equality only if $|Y(\omega)| = c|X(\omega)|$, where c is an arbitrary constant. Now recall that

$$\left|\int_a^b X(\omega)Y(\omega)\,d\omega\right| \leq \int_a^b |X(\omega)||Y(\omega)|\,d\omega = \int_a^b |X(\omega)Y(\omega)|\,d\omega \tag{B.4}$$

with equality if and only if $Y(\omega) = cX^*(\omega)$, where c is an arbitrary constant. Equation (B.2) immediately follows from Eqs. (B.3) and (B.4).

Appendix C

GRAM-SCHMIDT ORTHOGONALIZATION OF A VECTOR SET

We have defined the dimensionality of a vector space as equal to the maximum number of independent vectors in the space. Thus in an N-dimensional space, there can be no more than N vectors that are independent. Alternatively, it is always possible to find a set of N vectors that are independent. Once such a set is chosen, any vector in this space can be expressed in terms of (as a linear combination of) the vectors in this set. This set forms what we commonly refer to as a basis set, which forms the coordinate system. This set of N independent vectors is by no means unique. The reader is familiar with this fact in the physical space of three dimensions, where one can find an infinite number of independent sets of three vectors. This is clear from the fact that we have an infinite number of possible coordinate systems. An orthogonal set, however, is of special interest because it is easier to deal with as compared to nonorthogonal set. If we are given a set of N independent vectors, it is possible to obtain from this set another set of N independent vectors that is orthogonal. This is done by the Gram-Schmidt process of orthogonalization.

In the following derivation, we use the result [derived in Eq. (2.26)] that the projection (or component) of a vector $\boldsymbol{x}_2$ upon another vector $\boldsymbol{x}_1$ (see Fig. C.1) is $c_{12}\boldsymbol{x}_1$, where

$$c_{12} = \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{|\boldsymbol{x}_1|^2} y_1 \tag{C.1a}$$

The error in this approximation is the vector $\boldsymbol{x}_2 - c_{12}\boldsymbol{x}_1$, that is,

$$\text{Error vector} = \boldsymbol{x}_2 - \frac{\boldsymbol{x}_1 \cdot \boldsymbol{x}_2}{|\boldsymbol{x}_1|^2}\boldsymbol{x}_1 \tag{C.1b}$$

The error vector, shown dashed in Fig. C.1 is orthogonal to vector $\boldsymbol{x}_1$.

In order to get a physical insight into this procedure, we shall consider a simple case of 2-dimensional space. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be two independent vectors in a 2-dimensional space (Fig. C.1). We wish to generate a new set of two orthogonal vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ from $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. For convenience, we shall choose

$$\boldsymbol{y}_1 = \boldsymbol{x}_1 \tag{C.2}$$

We now find another vector $\boldsymbol{y}_2$ that is orthogonal to $\boldsymbol{y}_1$ (and $\boldsymbol{x}_1$. Figure C.1 shows that the error

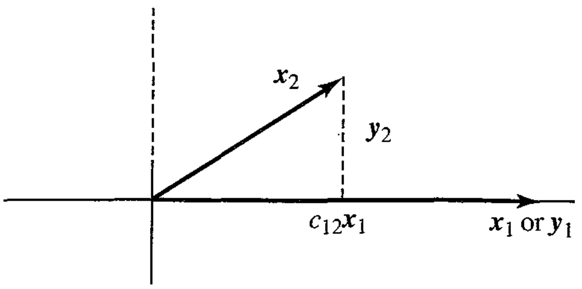


Figure C.1 Gram-Schmidt process for a 2-dimensional case.

vector in approximation of x_2 by y_1 (shown by dashed lines) is orthogonal to y_1, and can be taken as y_2. Hence,

$$\begin{aligned} y_2 &= x_2 - \frac{x_1 \cdot x_2}{|x_1|^2} x_1 \\ &= x_2 - \frac{y_1 \cdot x_2}{|y_1|^2} y_1 \end{aligned} \tag{C.3}$$

Equations (C.2) and (C.3) yield the desired orthogonal set. Note that this set is not unique. There is an inifinite number of possible orthogonal vector sets (y_1 , y_2) that can be generated from (x_1, x_2). In our derivation, we could as well have started with $y = x_2$ instead of $y_1 = x_1$. This starting point would have yielded an entirely different set.

The reader can extend these results to a 3-dimensional case. If vectors x_1, x_2, x_3 form an independent set in this space, then we form vectors y_1 and y_2 as in Eqs (C.2) and (C.3). To determine y_3, we approximate x_3 in terms of vectors y_1 and y_2. The error in this approximation must be orthogonal to both y_1 and y_2 and, hence, can be taken as the third orthogonal vector y_3. Hence,

$$\begin{aligned} y_3 &= x_3 - \text{sum of projections of } x_3 \text{ on } y_1 \text{ and } y_2 \\ &= x_3 - \frac{y_1 \cdot x_3}{|y_1|^2} y_1 - \frac{y_2 \cdot x_3}{|y_2|^2} y_2 \end{aligned} \tag{C.4}$$

These results can be extended to an N-dimensional space. In general, if we are given N independent vectors $x_1, x_2, \ldots, x_N$, then proceeding along similar lines, one can obtain an orthogonal set $y_1, y_2, \ldots, y_N$, where

$$y_1 = x_1$$

and

$$y_j = x_j - \sum_{k=1}^{j-1} \frac{y_k \cdot x_j}{|y_k|^2} y_k \qquad j = 2, 3, \ldots, N \tag{C.5}$$

Note that this is one of the infinitely many orthogonal sets that can be formed from the set x_1, x_2, ..., x_N. Moreover, this set is not an orthonormal set. The orthonormal set $\hat{y}_1$, $\hat{y}_2$, ..., $\hat{y}_N$ can be obtained by normalizing the lengths of the respective vectors,

$$\hat{y}_k = \frac{y_k}{|y_k|}$$

We can apply these concepts to signal space, because one-to-one correspondence exists between signals and vectors. If we have N independent signals $x_1(t),\ x_2(t),\ \ldots,\ x_N(t)$, we can form a set of N orthogonal signals $y_1(t),\ y_2(t),\ \ldots,\ y_N(t)$ as

$$\begin{aligned} y_1(t) &= x(t) \\ y_j(t) &= x_j(t) - \sum_{k=1}^{j-1} c_{kj} y_k(t) \qquad j = 2, 3, \ldots, N \end{aligned} \tag{C.6}$$

where

$$c_{kj} = \frac{\int y_k(t) x_j(t)\, dt}{\int y_k^2(t)\, dt} \tag{C.7}$$

Note that this is one of the infinitely many possible orthogonal sets that can be formed from the set $x_1(t),\ x_2(t),\ \ldots,\ x_N(t)$. The set can be normalized by dividing each signal $y_j(t)$ by its energy.

EXAMPLE C.1

The exponential signals

$$\begin{aligned} g_1(t) &= e^{-pt}u(t) \\ g_2(t) &= e^{-2pt}u(t) \\ &\vdots \\ g_N(t) &= e^{-Npt}u(t) \end{aligned}$$

form an independent set of signals in N-dimensional space, where N may be any integer. This set, however, is not orthogonal. We can use the Gram-Schmidt process to obtain an orthogonal set for this space. If $y_1(t),\ y_2(t),\ \ldots,\ y_N(t)$ is the desired orthogonal basis set, we choose

$$y_1(t) = g_1(t) = e^{-pt}u(t)$$

From Eqs. (C.6) and (C.7) we have

$$y_2(t) = x_2(t) - c_{12} y_1(t)$$

where

$$\begin{aligned} c_{12} &= \frac{\int_{-\infty}^{\infty} y_1(t) x_2(t)\, dt}{\int_{-\infty}^{\infty} y_1^2(t)\, dt} \\ &= \frac{\int_0^{\infty} e^{-pt} e^{-2pt}\, dt}{\int_0^{\infty} e^{-2pt}\, dt} \\ &= \frac{2}{3} \end{aligned}$$

Hence,

$$y_2(t) = (e^{-2pt} - \tfrac{2}{3} e^{-pt}) u(t)$$

Similarly, we can proceed to find the remaining functions $y_3(t),\ \ldots,\ y_N(t)$, and so on. The reader can verify that all this represents a mutually orthogonal set.

Appendix D
MISCELLANEOUS

D.1 L'Hôpital's Rule

If lim $f(x)/g(x)$ results in the indeterministic form 0/0 or ∞/∞, then

$$\lim \frac{f(x)}{g(x)} = \lim \frac{\dot{f}(x)}{\dot{g}(x)}$$

D.2 Taylor and Maclaurin Series

$$f(x) = f(a) + \frac{(x-a)}{1!}\dot{f}(a) + \frac{(x-a)^2}{2!}\ddot{f}(a) + \cdots$$

$$f(x) = f(0) + \frac{x}{1!}\dot{f}(0) + \frac{x^2}{2!}\ddot{f}(0) + \cdots$$

D.3 Power Series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \cdots \qquad x^2 < \frac{\pi^2}{4}$$

$$Q(x) = \frac{e^{-x^2/2}}{x\sqrt{2\pi}}\left(1 - \frac{1}{x^2} + \frac{1 \cdot 3}{x^4} - \frac{1 \cdot 3 \cdot 5}{x^6} + \cdots\right)$$

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots + \binom{n}{k}x^k + \cdots + x^n$$

$$\approx 1 + nx \qquad |x| \ll 1$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \qquad |x| < 1$$

D.4 Sums

$$\sum_{m=0}^{k} r^m = \frac{r^{k+1}-1}{r-1} \qquad r \neq 1$$

$$\sum_{m=M}^{N} r^m = \frac{r^{N+1}-r^M}{r-1} \qquad r \neq 1$$

$$\sum_{m=0}^{k} \left(\frac{a}{b}\right)^m = \frac{a^{k+1}-b^{k+1}}{b^k(a-b)} \qquad a \neq b$$

D.5 Complex Numbers

$$e^{\pm j\pi/2} = \pm j$$

$$e^{\pm jn\pi} = \begin{cases} 1 & n \text{ even} \\ -1 & n \text{ odd} \end{cases}$$

$$e^{\pm j\theta} = \cos\theta \pm j\sin\theta$$

$$a + jb = re^{j\theta} \qquad r = \sqrt{a^2+b^2}, \quad \theta = \tan^{-1}\left(\frac{b}{a}\right)$$

$$(re^{j\theta})^k = r^k e^{jk\theta}$$

$$(r_1 e^{j\theta_1})(r_2 e^{j\theta_2}) = r_1 r_2 e^{j(\theta_1+\theta_2)}$$

D.6 Trigonometric Identities

$$e^{\pm jx} = \cos x \pm j\sin x$$

$$\cos x = \frac{1}{2}(e^{jx} + e^{-jx})$$

$$\sin x = \frac{1}{2j}(e^{jx} - e^{-jx})$$

$$\cos\left(x \pm \frac{\pi}{2}\right) = \mp \sin x$$

$$\sin\left(x \pm \frac{\pi}{2}\right) = \pm \cos x$$

$$2 \sin x \cos x = \sin 2x$$

$$\sin^2 x + \cos^2 x = 1$$

$$\cos^2 x - \sin^2 x = \cos 2x$$

$$\cos^2 x = \frac{1}{2}(1 + \cos 2x)$$

$$\sin^2 x = \frac{1}{2}(1 - \cos 2x)$$

$$\cos^3 x = \frac{1}{4}(3 \cos x + \cos 3x)$$

$$\sin^3 x = \frac{1}{4}(3 \sin x - \sin 3x)$$

$$\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$$

$$\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y$$

$$\tan(x \pm y) = \frac{\tan x \pm \tan y}{1 \mp \tan x \tan y}$$

$$\sin x \sin y = \frac{1}{2}[\cos(x - y) - \cos(x + y)]$$

$$\cos x \cos y = \frac{1}{2}[\cos(x - y) + \cos(x + y)]$$

$$\sin x \cos y = \frac{1}{2}[\sin(x - y) + \sin(x + y)]$$

$$a \cos x + b \sin x = C \cos(x + \theta)$$

in which $C = \sqrt{a^2 + b^2}$ and $\theta = \tan^{-1}\left(\frac{-b}{a}\right)$

D.7 Indefinite Integrals

$$\int u\, dv = uv - \int v\, du$$

$$\int f(x)\dot{g}(x)\, dx = f(x)g(x) - \int \dot{f}(x)g(x)\, dx$$

$$\int \sin ax\, dx = -\frac{1}{a} \cos ax \qquad \int \cos ax\, dx = \frac{1}{a} \sin ax$$

$$\int \sin^2 ax\, dx = \frac{x}{2} - \frac{\sin 2ax}{4a} \qquad \int \cos^2 ax\, dx = \frac{x}{2} + \frac{\sin 2ax}{4a}$$

$$\int x \sin ax\, dx = \frac{1}{a^2}(\sin ax - ax \cos ax)$$

$$\int x\cos ax\,dx = \frac{1}{a^2}(\cos ax + ax\sin ax)$$

$$\int x^2\sin ax\,dx = \frac{1}{a^3}(2ax\sin ax + 2\cos ax - a^2x^2\cos ax)$$

$$\int x^2\cos ax\,dx = \frac{1}{a^3}(2ax\cos ax - 2\sin ax + a^2x^2\sin ax)$$

$$\int \sin ax\sin bx\,dx = \frac{\sin(a-b)x}{2(a-b)} - \frac{\sin(a+b)x}{2(a+b)} \qquad a^2 \neq b^2$$

$$\int \sin ax\cos bx\,dx = -\left[\frac{\cos(a-b)x}{2(a-b)} + \frac{\cos(a+b)x}{2(a+b)}\right] \qquad a^2 \neq b^2$$

$$\int \cos ax\cos bx\,dx = \frac{\sin(a-b)x}{2(a-b)} + \frac{\sin(a+b)x}{2(a+b)} \qquad a^2 \neq b^2$$

$$\int e^{ax}\,dx = \frac{1}{a}e^{ax}$$

$$\int xe^{ax}\,dx = \frac{e^{ax}}{a^2}(ax-1)$$

$$\int x^2e^{ax}\,dx = \frac{e^{ax}}{a^3}(a^2x^2 - 2ax + 2)$$

$$\int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2+b^2}(a\sin bx - b\cos bx)$$

$$\int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2+b^2}(a\cos bx + b\sin bx)$$

$$\int \frac{1}{x^2+a^2}\,dx = \frac{1}{a}\tan^{-1}\frac{x}{a}$$

$$\int \frac{x}{x^2+a^2}\,dx = \frac{1}{2}\ln(x^2+a^2)$$

INDEX

2-binary 1-quaternary (2B1Q), 371
Aliasing error (spectral folding), 60, 133, 257
Alternate mark inversion (AMI), 295, 306, 359
American national standards institute (ANSI), 432
Amplitude modulation (AM), 162–170, 537–541, 718
Amplitude phase shift keying (APK), 341
Amplitude shift keying, 337–339, 591, 595
Amplitude spectrum, 49, 74
Analog-to-digital conversion, 5, 265
Analog message, 3
Analog signals, 20
  conversion to digital, 5, 265–270
Angle modulation, 208, 209, 541
Asymmetric digital subscriber line (ADSL), 393
Asynchronous transfer mode (ATM), 383
Autocorrelation function, 40, 125, 127, 490
Avalanche photodiode (APD), 423

Bandpass limiter, 234
Bandpass signals, 95, 514
  distortionless transmission, 102
Bandwidth, 8, 81
  and duration of signal, 89
  essential, 117
  of product of signals, 89
  of rect function, 81
Baseband signal, 1, 151
Basic rate interface (BRI), 369
Bayes' receiver, 670
Bayes' rule, 439
Bearer channel (B channel), 369
Bipolar return to zero (BPRZ), 359
Binary
  with 8 zero substitution (B8ZS), 309, 360
  with 6 zero substitution (B6ZS), 309
Binary 8 zero suppression (B8ZS), 309, 360
Binary message, 3, 20
Binary symmetric channel (BSC) 447
  channel capacity of, 696
Bipolar (pseudoternary) signaling, 295, 306
Bit (binary digit), 263
Bit error rate (BER), 584, 609, 613, 617
Bit stuffing, 345
Block codes, 728, 731
Bose-Chaudhuri-Hocquenghen (BCH) codes, 745
Broadband ISDN (BISDN), 390
Butterworth filters, 108
Burst-error-detection and -correction codes, 745

Capture effect, 242
Capture range of PLL, 186
Carrier, 10
Carrier acquisition, 183
Carrier communication, 151
Carson's rule, 219
Causal systems, 107
CCITT, 275
Cellular and digital packet data (CDPD), 412
Cellular telephone, 404, 412
Central-limit theorem, 472
Channel, 1
Channel bank, 358
Channel capacity, 693, 709

Channel matrix, 694
Channel service unit (CSU), 363
Chebyshev's inequality, 471
Chip, 407
Chip rate, 407
Circuit, switched, 368
Code-division multiple access (CDMA), 406, 412, 600
Code efficiency, 687, 728
Code polynomial, 738
Code rate, 728
Code tree, 748
Coefficient of correlation, 36, 476, 657
Coherent detection, 154, 590
Communication systems, 1
  analog, 532
  digital, 294, 577
Compact codes, 685
Companding, 268, 563
Compressors, 268, 563
Conditional probability, 439
Conjugate symmetry property, 75
Consultative Committee on International Telegraphy and Telephony (CCITT), 275
Continuous-time signals, 20
Convolutional codes, 728, 747
Correlation, 35, 473–476
Correlation coefficient, 36
Correlation functions, 39
Correlator detector, 581
Costas loop, 187
Cross-correlation function, 40, 509
Cross-power spectral density, 510
Cumulative distribution function (CDF), 449
Cyclic codes, 737–746
Cyclostationary random process, 504n

D4 framing (DF), 355
Data channel (D channel), 369
Decision regions, 646
Deemphasis, 243
Delta modulation, 281–288
  adaptive, 285
  comparison with PCM, 287
  double integration, 285
  output SNR, 286
  overloading in, 284–285
  threshold of coding, 284
Demodulation, 10, 154, 161, 167, 177, 233
Detection error probability, 329, 586–589
Deviation ratio, 219
Differential coding, 318, 600
Differential GPS, 416
Differential pulse code modulation (DPCM), 278–281
Differentially coherent phase-shift keying (DPSK), 599
Digit interleaving, 343
Digital carrier systems, 337, 590
Digital communication systems, 4, 294, 577
  advantages of, 263
Digital filters, 109
Digital hierarchy, 346–348
Digital multiplexing, 342
Digital service unit (DSU), 363
Digital signal cross connect, level 1 (DSX1), 363
Digital signal at level one (DS1), 355
Digital signal at level zero (DS0), 354
Digital signals, 3, 20
  noise immunity of, 4, 263
Direchlet conditions, 49
  strong, 49
  weak, 49
Direct sequence spread spectrum (DS/SS), 407, 602
Discrete cosine transform (DCT), 396
Discrete Fourier transform, 60, 130
Discrete multitone (DMT), 394
Discrete-time signals, 20
Dispersion
  material, 425
  multimode, 425
  profile, 425
  waveguide, 425
Distortion, 110
  in audio and video signals, 104
  due to multipaths, 114
  linear, 110
  nonlinear, 111
Distortionless transmission, 102
Double sideband modulation (DSB), 152, 534
Duality, time frequency, 86
Duobinary signaling, 295, 317, 527
  modified, 296n

Elastic store, 345
Energy, 15
  of complex signals, 15
  of modulated signals, 120
Energy spectral density (ESD), 116, 121
Ensemble of a random process, 487
Entropy of a source, 682
Envelope detection, 168, 233, 538
Envelope (group) delay, 104n
Equalizer, 323–325
  automatic and adaptive, 326

  least mean square error (LMSE), 326
  zero-forcing, 323
- Equivalent signal set, 662
- Equivocation, 694
- Error-free communication, 690, 699, 712
- Essential bandwidth, 117, 134
- Estimation, linear mean square, 476
- Events, 435
  - compliment of, 435
  - disjoint, 436
  - independent, 439
  - intersection of, 436
  - joint, 436
  - mutually exclusive, 436
  - union of, 435
- Extended superframe (ESF), 277, 356

- Fading, 115
  - selective, 115
- Fast Fourier transform, 61, 135
- Feedback decoding, 755
- Fiber-to-the-curb (FTTC), 393
- Fiber-optic link around the globe (FLAG), 424
- First-order hold circuit, 134, 255
- Folding frequency, 257
- *Fortune* magazine, 226
- Fourier series, 43
  - convergence at jump discontinuities, 49
  - effect of symmetry, 53
  - exponential, 53
  - generalized, 43
  - periodicity of, 47
  - trigonometric, 44
  - trigonometric, compact, 45
- Fourier transform, 74
  - conjugate symmetry property of, 75
  - direct, 74
  - discrete, 60
  - duality (symmetry) property, 86
  - existence of, 76
  - fast, 61
  - frequency-convolution property, 98
  - frequency-shifting property, 92
  - inverse, 74
  - linearity of, 76
  - scaling property, 88
  - symmetry (duality) property, 86
  - time-convolution property, 98
  - time-differentiation property, 99
  - time-integration property, 99
  - time-shifting property, 90
- Frame, 355
- Frame relay, 372
- Frequency
  - folding, 257
  - fundamental, 44
  - instantaneous, 208, 209
  - *n*th harmonic, 44
- Frequency compression feedback (FCF), 555
- Frequency converter (mixer), 160, 189
- Frequency-convolution property, 98
- Frequency discriminators, 236
- Frequency-division multiple access (FDMA), 410
- Frequency-division mutiplexing, 12, 189
- Frequency domain description of a signal, 49
- Frequency-hopping spread spectrum system (FH/SS), 411, 606
  - fast hopping, 411
  - slow hopping, 411
- Frequency modulation (FM), 10, 211
- Frequency-shift keying (FSK), 337, 592, 598
- Full cosine roll-off characteristic, 314

- Gate function, 78
- Gaussian probability density, 331, 452
- Gaussian random process, 632
  - bandpass, 519
  - transmission through linear systems, 634
- Generalized angle, 209
- Generalized function, 29
- Generator matrix, 731
- Generator polynomial, 738
- Global positioning system (GPS), 413
- Global system for mobile communications (GSM), 412
- Graded index fiber, 425
- Gram-Schmidt orthogonalization process, 632, 768
- Ground reflected wave, 419
- Ground wave, 419

- Hamming bound, 729
- Hamming codes, 730
- Hamming distance, 691, 729
- Hamming sphere, 729
- High definition television (HDTV), 400
- High-bit rate digital subscriber line (HDSL), 364
- High-density bipolar (HDB) signaling, 308
- Hilbert transform, 173
- Hold-in range of PLL, 186
- Hybrid circuit, 427
- Hybrid fiber coax (HFC), 393

- Ideal filters, 106
- Image station, 190

Impulse function, 28
Industrial, scientific, and medical frequency band (ISM), 412
Information measure, 680
Instantaneous frequency, 208–210
Integrated services digital network (ISDN), 369
Interference
  in AM systems, 242n
  in angle-modulated systems, 241
Interleaving effect, 247
Intermediate frequency (IF), 189–190, 245
Interpolation (sinc) function, 80
Intersymbol interference (ISI), 310

Jitter, 202, 328

Laplace probability density function, 565
Lasers, 423
Light emitting diodes (LED), 423
Line coding, 295, 297
Linear mean-square estimation, 476
Linear predictor (estimator), 276, 476
Local access and transport areas (LATA), 432
Local exchange companies (LEC), 432
Lock range of PLL, 186
Logarithmic units, 273
Low earth orbit (LEO), 405

Manchester (split-phase) signaling, 304
M-ary communication, 334, 608
M-ary message, 4, 334, 608
Matched filter, 580
Maximum a posteriori probability (MAP) detector, 643
Maximum frequency deviation, 217
Maximum length shift register sequences, 601
Maximum-likelihood decoding, 734, 750
Maximum-likelihood receiver, 671
Mean square bandwidth, 549
Minimax receiver, 672
Minimum energy signal set, 664
Minimum weight vector, 734
Mobile radio, 404
Mobile telephone switching office (MTSO), 404
Modem, 342
Modified duobinary, 296n
Modulation, 10, 93
  amplitude (AM), 93, 162, 537, 718
  double sideband (DSB), 152, 534
  frequency (FM), 10, 209
  phase, (PM), 10, 209
  single sideband (SSB), 171, 535
  vestigial sideband (VSB), 179
Modulation index, 164, 219
Modulators
  balanced, 157, 159
  FM, 229
  nonlinear, 156
  switching, 157
Moments of random variables, 466
Motion picture experts group (MPEG), 396
Mulitamplitude signaling, 335, 608, 613
Multipath effect, 114, 605
Multiphase signaling (MPSK), 611
Multiple access, 410, 603
Multiple access interference (MAI), 603
Multiplexing
  asynchronous channels, 345
  digital, 342
  frequency division, 12, 189
  time division, 12, 261
Multitone signaling, 615
Mutual information, 695

Narrowband angle modulation, 216, 548
Natural binary code (NBC), 262
Near-far problem, 410, 604
Negative frequency, 57
Noise, 3
Noncoherent detection, 594, 618
Nonreturn to zero (NRZ), pulses, 296
North American digital hierarchy, 346, 365
Null event, 435
Nyquist criteria for zero ISI, 310
Nyquist interval, 253
Nyquist sampling rate, 253

On-off keying (OOK), 337
On-off signaling, 295, 304, 587
Optical carrier (OC-n), 378
Optimum filters, 522, 580
Optimum preemphasis-deemphasis, 567
Optimum receiver, 626
Optical communication systems, 422
Orthogonal set, 42
Orthogonal signals, 33, 24
Orthogonal signaling, 588, 615, 658
Orthogonal vectors, 32
Orthogonality
  of exponential set, 765
  of trigonometric set, 764
Orthonormal set, 42

Packet switched, 368

Paley-Wiener criterion, 107, 107n
Partial response signals, 316
Parity-check digits, 731
Parity-check matrix, 733
Parseval's theorem, 59, 115
PCM, 6, 262, 557
PCM repeater, 442
Permanent virtual circuit (PVC), 373
Phase delay, 104n
Phase-lock loop (PLL), 184, 236, 557
  capture (pull-in) range, 186
  hold-in (lock) range, 186
Phase modulation (PM), 10, 209, 219
Phase-shift keying (PSK), 337, 339, 591
Phase spectrum, 49, 75
Plesiochronous, 367
Pointer, 381
Polar signaling, 295, 302, 586
Power
  of complex signals, 15
  of modulated signals, 130
  of real signals, 15
Power spectral density, 123, 496
  interpretation of, 126
Prediction coefficients, 276
Preemphasis-deemphasis (PDE), 243, 567
Preenvelope, 172n
Primary rate interface (PRI), 369
Probability density function (PDF), 331, 450
  conditional, 439, 461
  joint, 459
Pseudonoise (PN) sequence, 407
  generation of, 600
Pseudoternary (bipolar) signaling, 295
Public switched telephone network (PSTN), 430
Pull-in range of PLL, 186
Pulse-amplitude modulation (PAM), 260
Pulse-code modulation (PCM), 6, 262, 557
  companded, 268, 563
  noise due to detection errors, 468–469
  output SNR, 272, 559, 565
  of quantization noise, 266, 468, 565
  synchronizing and signaling, 276
Pulse-position modulation (PPM), 260
Pulse shaping, 310
Pulse stuffing, 345
Pulse-width (or duration) modulation (PWM), 260

Quadrature-amplitude modulation (QAM), 170, 341, 613, 652
Quadrature multiplexing, 171
Qualcomm code-excited linear prediction (QCELP), 413
Quantization, 5, 262
Quantization noise, 265, 467, 565

Raised cosine characteristics, 314
Rake receiver, 410, 606
Random processes, 487–495
  bandpass, 514
  classification of, 492
  cross-power spectral density, 510
  cyclostationary, 504n
  ergodic, 494
  gaussian, 519, 633
  geometrical representation of, 635
  incoherent, 509
  independent, 509
  mean square value (power) of, 499
  orthogonal, 509
  power spectral density (PSD), 499
  stationary, 492
  uncorrelated, 509
  wide-sense (weakly) stationary, 493
Random variables, 445
  continuous, 446
  correlation of, 473
  discrete, 445
  means, 463
Rate of communication, 8, 696, 709, 711
Ratio detector, 236
Rayleigh probability density, 461, 595
Realizable systems, 107
Rectifier detector, 167
Redundancy, 13, 687
Regeneration, 358
Regenerative repeater, 4, 263, 322
Relative frequency, 436
Return to zero (RZ) pulses, 296
Rice probability density, 521, 595
Robbed bit signaling, 277, 356
Roll-off factor, 314

Sample function, 487
Sample point, 435
Sample space, 435
Sampling, instantaneous, 251
  practical, 258
Sampling (sifting) property, 28
Sampling theorem, 5, 251
Satellite communication systems, 421
Scaling property of the Fourier transform, 88
Scatter diagram, 474
Schwartz inequality, 36, 568, 578, 766
Scrambling, 319
Sequential decoding, 752

Sidebands, 152
Signal classification, 20
  aperiodic, 20, 21
  analog, 20
  continuous-time, 20
  deterministic, 24, 434
  digital, 20
  discrete-time, 20
  energy, 20, 23
  periodic, 20, 21
  power, 20, 23
  random, 20, 24, 434
Signal comparison (correlation), 35
Signal component, 32
Signal distortion
  linear, 110
  multipath effect, 114
  nonlinear, 111
Signal energy, 15, 23
Signal power, 15, 23
Signal size, 14
Signal-to-noise ratio (SNR), 3, 8, 14
  in amplitude-modulated systems, 534–541
  in angle-modulated systems, 541–552
  in baseband systems, 532
  in DM, 286
  exchange with bandwidth, 12, 611, 619, 718
  in PCM, 272
Signal space, 626, 642
Signaling system 7 (SS7), 383
Signals
  analog, 3, 20
  bandpass, 95
  digital, 3, 20
  energy, 15
  periodic, 21
  power, 15
  vector representation of, 628
Sign (sgn) function, 84
Simplex signal set, 668
Sinc function, 80
Single sideband modulation (SSB), 171–179, 535
Sky wave, 419
Slope detector, 236
Source encoding, 684
Spectrum
  amplitude, 49, 75
  continuous, 75
  discrete (or line), 49
  energy density, 116
  exponential Fourier, 55
  phase, 49, 75
  power density, 123
Split-phase (twinned binary) signaling, 304
Spread spectrum systems, 406
  direct sequence (DS/SS), 407, 602
  frequency-hopping, (FH/SS), 411, 606
Square-law detector, 205
Standard deviation, 467
Statistical time division multiplexing (STATDM), 383
Stochastic processes, 487
Superframe, 277, 355
  extended, 277, 357
Superheterodyne receiver, 189
Switched circuit, 368
Switched multi-megabit data service (SMDS), 391
Symmetry (duality) property, 86
Synchronization, 276, 328, 622
Synchronous detection or demodulation, 161
Synchronous optical network (SONET), 379
Synchronous payload envelope (SPE), 379
Synchronous transport signal (STS-n), 378
Synchronous satellite, 421
Syndrome, 733
Systematic codes, 731, 740

T-1 carrier system, 274–278, 343, 346, 347, 355
Television, 191–201
Thermal noise, 511, 528
Threshold detection, 39, 455, 579
Threshold effect, 540, 553, 560
Time assignment system interpolation, 430
Time convolution property, 98
Time correlation, 121, 125
Time-differentiation property, 99
Time-division multiple-access (TDMA) systems, 345, 410, 413
Time-division multiplexing, 12, 261
Time domain description of a signal, 49
Time-integration property, 99
Time-shifting property, 90
Timing extraction, 328
Timing jitter, 328, 329
Tone modulation, 164, 220
Transmission media, 416–427
Transparency, 297
Transversal filter, 115, 148, 279, 319
Trellis diagram, 751
Triangle function, 79
Tropospheric scattering, 421
Tropospheric wave, 419
Trunk circuit, 432
Twinned binary (split-phase or Manchester) signaling, 304

Unrealizable systems, 107

Variance, 467
Vector space, 41, 626–628
Very high data rate ADSL (VDSL), 395
Vestigal sideband modulation, 179–182
Vestigal spectrum, 314
Video-on-demand (VOD), 392
Virtual channel, (VC), 384
Virtual path (VP), 384
Virtual tributary (VT), 379
Viterbi algorithm, 750
Voltage-controlled oscillator (VCO), 185, 236, 555

Weaver's method of SSB generation, 176
Wiener-Hopf filter, 522
Wiener-Khinchine relation, 499

x-DSL, 393

Zero padding, 132
Zero-forcing equalizer, 323